Vom Genom zum Modell

Qualitative Verbesserung der Funktionsvorhersage von Enzymen und Abbildung der Informationen auf einem Stoffwechselnetzwerk

Von der Fakultät für Lebenswissenschaften

der Technischen Universität Carolo-Wilhelmina

zu Braunschweig

zur Erlangung des Grades einer

Doktorin der Naturwissenschaften

(Dr. rer. nat.)

genehmigte

D i s s e r t a t i o n

von Susanne Elisabeth Quester

aus Köln

1. Referent: Prof. Dr. Dietmar Schomburg

2. Referent: Prof. Dr. Dieter Jahn

eingereicht am: 16.11.2011

mündliche Prüfung (Disputation) am: 14.03.2012

Druckjahr 2012

Vorveröffentlichungen der Dissertation

Teilergebnisse aus dieser Arbeit wurden mit Genehmigung der Fakultät für Lebenswissenschaften, vertreten durch den Mentor der Arbeit, in folgenden Beiträgen vorab veröffentlicht:

Publikationen

Quester, S. & Schomburg, D.. EnzymeDetector: an integrated enzyme function prediction tool and database. *BMC Bioinformatics* 2011, **12:**376.

Tagungsbeiträge

Quester S. & Schomburg, D.: EnzymeDetector: a comparison of genome annotations from common databases and an up-to-date BLAST-Search to help the user find the right annotation (Poster) German Conference of Bioinformatics (2009)

Quester S. & Schomburg, D.: EnzymeDetector: integrated enzyme function prediction tool (Poster) German Conference of Bioinformatics (2010)

Quester S. & Schomburg, D.: Enzyme Detector – integrated genomic enzyme function prediction (Poster) German Conference of Bioinformatics (2011)

Hiermit möchte ich meiner Familie und meinen Freunden danken, die mich immer bestärkt und unterstützt haben.

Besonders danke ich Tomek, der mich seelisch immer wieder aufgebaut hat und sich in den stressigsten Phasen mit unendlich vielen Apfel-Crumbles und belgischen Waffeln um mich gekümmert hat! ;D

Auch der Arbeitsgruppe Schomburg möchte ich dafür danken, dass alle immer als Ansprechpartner zur Verfügung standen. Adam danke ich für die viele Unterstützung bei der Entwicklung der Website

Besonders möchte ich Alex und meiner Mutter danken, der sich die Mühe gemacht haben, diese gesamte Arbeit genauestens Korrektur zu lesen!

Und natürlich danke ich Prof. Schomburg, dass er mir diese Arbeit ermöglicht hat und mit vielen Ideen und Vorschlägen zur Seite stand, wenn ich mal nicht weiter gekommen bin.

Ebenfalls danken möchte ich Prof. Jahn, der sich bereit erklärt hat als Zweigutachter diese Arbeit zu bewerten.

Inhaltsverzeichnis

Zusammenfassung

Das relativ neue Teilgebiet der Biowissenschaften, die Systembiologie, versucht Organismen in ihrer Gesamtheit zu verstehen. Hierfür werden Modelle von Organismen erstellt und anschließend mit mathematischen Methoden analysiert und untersucht.

Um das metabolische Verhalten eines Organismus simulieren zu können, ist es essentiell, einen umfassenden Überblick über seinen Vorrat an Enzymen zu haben. Hierfür kann man auf die Genomannotationen von entsprechenden Datenbanken, wie NCBI, KEGG und PEDANT zurückgreifen. Es hat sich allerdings herausgestellt, dass die Annotationen dieser Datenbanken nur in einem Viertel der Fälle identisch sind. Somit ergibt sich die Schwierigkeit zu entscheiden, welche der Informationen als richtig angesehen wird.

Auch konnte gezeigt werden, dass nur mit den kombinierten Daten mehrerer Annotationsquellen eine umfassende Informationsabdeckung erreicht und der Erwartungswert von einem Drittel der enzymbildenden Gene pro Genom erzielt werden konnte. Dieser Wert leitet sich vom sehr gut untersuchten Organismus *E. coli* ab. Hierbei lieferte die durchgeführte BLAST-basierte Genomannotation einen Großteil der Informationen. Ein Drittel dieser gefundenen Annotationen hatte sogar einen E-value von unter 10^{-120}. Um ein umfassendes Bild des Enzymvorrats eines Organismus zu erhalten, wurde das Programm EnzymeDetector entwickelt. Es liefert eine integrierte Genomannotation, welche den aktuellen Wissensstand repräsentiert. Die zusammengeführten Annotationsquellen werden relativ zueinander gewertet basierend auf einer Auswertung der Quellen gegen einen Wahrheitswert. Nur so ist es möglich zu entscheiden, welche der gefundenen Annotationen für ein Gen die richtige ist.

Die Daten aus dem EnzymeDetector können dann auf der mitentwickelten Stoffwechselkarte visualisiert werden. Dies erleichtert die Auswertung der Daten und das Auffinden von Lücken im Stoffwechsel im Zuge der Modellierung. Die Lücken können mit Hilfe des Programms gapFiller geschlossen werden. Unterstützt wird der gapFiller durch den erweiterten AtomMapper. Dieser ermöglicht das Verfolgen von Atomen durch ganze Reaktionen und ist eine Entscheidungshilfe um festzulegen, welche biologisch sinnvollen Pfade beschritten werden dürfen ausgehend von einem Metabolit.

Durch den Zusammenschluss der vorgestellten Programme wird das Erstellen von Modellen basierend auf der Genomannotation erleichtert. Es können Schritte automatisiert ausgeführt werden, die vorher zeitaufwendig manuell erledigt werden mussten. Die Integration der Vielzahl an Datenquellen wäre manuell kaum realisierbar. Auch wenn weiterhin noch eine manuelle Erweiterung der Genomannotation vonnöten ist, so bieten die Programme doch eine große Hilfe.

Summary

Systems biology is a relatively new area of the life sciences that tries to understand organisms as a whole. For this goal, models of organisms are built and subsequently analysed using mathematical algorithms.

But to simulate the metabolic behaviour of an organism, it is crucial to have a comprehensive overview of its enzyme pool. Genome annotations from appropriate databases such as NCBI, KEGG, and PEDANT can be used for this purpose. However, the genome annotations from those databases have been found to be indentical in only 25 % of all cases. This leads to the difficult task of determining which annotation is correct.

Furthermore, it could be shown that only by using data integrated from several annotation sources, the expected value of one-third enzyme producing genes per genome could be achieved. This expected value is deflected from the rate of enzyme producing genes of the well-analysed organism *E. coli*. The BLAST-based genome annotation that was performed in addition to database searches provided a good portion of this information. One-third of the annotations that were found had an E-value below 10^{-120}.

The program EnzymeDetector was developed in order to gain a comprehensive picture of the enzyme pool of an organism. It provides an integrated genome annotation that represents the up-to-date level of knowledge. The integrated annotation sources are evaluated in relation to one another, based on an evaluation of the sources against a standard of truth. Only then it is possible to determine, which annotation is correct.

The data from the program EnzymeDetector can be visualised on the co-developed pathway map. This map facilitates the interpretation of the data and the identification of pathway gaps. Those gaps can be filled by the program gapFiller. The gapFiller is supported by the extended version of the program AtomMapper. In this version, the AtomMapper allows tracing atoms trough reactions. Thereby it can be used to guide the decision, which paths can be taken starting from one metabolite and which are not admissible.

The combination of these programs makes the development of models based on genome annotation much easier. Steps can be performed automatically that before had to be done in a time consuming process by hand. Even if a manual refinement of the genome annotation is necessary, the programs provide a basic model that can be worked with and thereby are a big help.

1 Einleitung

Für die Rekonstruktion metabolischer Netzwerke ist eine verlässliche und möglichst umfassende Genomannotation des zu modellierenden Organismus essentiell. Dank der dramatisch gewachsenen Sequenzierungsgeschwindigkeit, mit der die Annotationsgeschwindigkeit im Vergleich nicht mithalten kann, stehen heute riesige Datenmengen zur Verfügung. Hochdurchsatzverfahren sorgen dafür, dass die Zahl der sequenzierten Genome, Metabolom- und Proteomanalysen rapide wächst. Die Menge der Daten und deren hohe Vernetzung machen es unmöglich, diese Daten per Hand auszuwerten. Auf Grund dessen hat sich das Feld der Bioinformatik entwickelt. Sie beschäftigt sich sowohl mit der Speicherung und Integration der Daten als auch mit deren Auswertung und Interpretation.

1.1 Biologische Datenbanken

Für die Speicherung der verschiedenen biologischen Daten wurden unterschiedliche Datenbanken erstellt. Diese sind meist in Form von Datenbanksystemen, d.h. relationalen Datenbanken, organisiert.

Sequenzdatenbanken beinhalten die grundlegendsten biologischen Daten, die Sequenzen. Die bekanntesten Vertreter hiervon sind die GenBank[1] die Nukleotidsequenzen enthält, und die UniProt-Datenbank[2], die Proteinsequenzen beinhaltet. Es existiert eine Fülle weiterer Datenbanken für die verschiedensten biologischen Fragestellungen.

1.2 -omiks

Durch ein modernes Teilgebiet der Biologie, den -omiks, werden auf Grund von Hochdurchsatzverfahren riesige Datenmengen produziert. Biologische Datenbanken spielen eine entscheidende Rolle in der Speicherung dieser Daten. Benannt werden die verschiedenen Teilgebiete der -omiks entsprechend den Bereichen des Organismus, die untersucht werden.

Das Gebiet der Genomik behandelt das Genom bzw. die Gesamtheit aller Gene eines Organismus und deren Wechselwirkungen. Bei den Genomik-Daten handelt es sich um statische Daten, d.h. sie ändern sich über die Zeit nicht.

Die Daten der anderen -omik-Teilgebiete liefern den Zustand eines Organismus zu einem bestimmten Zeitpunkt. Dieser kann sich sehr schnell ändern, zum Beispiel, wenn eine Stressantwort ausgelöst wird oder ähnliches.

Proteomik beschäftigt sich mit dem Proteom einer Zelle bzw. eines Organismus, d.h. der Gesamtheit aller Proteine zu einem bestimmten Zeitpunkt. Proteine haben verschiedene Rollen im Organismus. Enzyme sind für den Stoffumsatz verantwortlich, Transportproteine für den Transport wichtiger chemischer Verbindungen (wie zum Beispiel den Sauerstofftransport im Blut) und Membranrezeptoren für die Weiterleitung von Signalen. Auch für die Strukturgebung sind die Proteine eine wichtige Stoffklasse. Auf Grund dieser vielfältigen Funktionalität der Proteine ist es wichtig, ihre Zusammensetzung zu kennen und die Interaktionen zwischen den einzelnen Proteinen zu untersuchen.

Im Gegensatz zur Proteomik beschäftigt sich die Metabolomik mit der Identifizierung und Quantifizierung aller Metabolite einer Zelle. Bei Metaboliten handelt es sich um die Zwischenprodukte von Stoffwechselvorgängen. Sie spielen eine wichtige Rolle in der Regulierung von Stoffkreisläufen und Zellaktivitäten. Auch bilden sie Grundbausteine der Strukturelemente (z.B. Zellwandbestandteile und Membranlipide), sind Energieträger, -überträger- und -speicher und können als Kommunikationsmittel fungieren (z.B. Farb- und Duftstoffe) uvm.

1.3 Systembiologie

Die -omik-Untersuchungen sind wichtiger Bestandteil der Systembiologie. Dieser Teilbereich der Biologie versucht, Organismen in ihrer Gesamtheit zu verstehen und alle dynamischen Prozesse auf allen Ebenen abzubilden[3].

Hierfür gibt es zwei Ansätze: Beim Bottom-up-Ansatz ist es erforderlich, dass ein Großteil der Gene und deren regulative Beziehungen untereinander bekannt sind. Der Top-down-Ansatz versucht, Daten aus Hochdurchsatzverfahren in Beziehung zu setzen.

Die Funktionsweise des Organismensystems soll durch Betrachtung der einzelnen Komponenten verstanden werden. Diese Nachbildung der Wirklichkeit bietet den großen Vorteil, dass mit einem erstellten Modell Simulationen und Vorhersagen möglich sind, ohne dass diese im Labor durchgeführt und getestet werden müssen. So ist die Suche nach wirtschaftlichen Mutanten eines Organismus schneller und effizienter als im Labor. Auch die Vorhersage von Nebenwirkungen in der Pharmaindustrie kann so optimiert werden.

1.4 Biologische Modelle

Es gibt verschiedene Ebenen, auf denen biologische Modelle erstellt werden können, abhängig von den Daten, die in Zusammenhang gebracht werden. Je größer die Menge und Vielfalt an Daten, die zur Verfügung stehen, desto wirklichkeitsgetreuer kann ein Modell erstellt werden.

Modelliert werden u.a. Stoffwechselnetzwerke, Signaltransduktionsnetzwerke und transkriptionsregulierende Netzwerke. Ein umfassendes Bild eines Organismus kann nur durch ein integriertes Netzwerk all dieser Netzwerktypen erhalten werden. Nur so wäre es möglich, eine Simulation eines Organismus zu erstellen, die der Wirklichkeit entspricht. Da dies aber sehr komplex ist und meist die entsprechenden Daten, die nötig wären, nicht vorliegen, werden diese Netzwerktypen oft unabhängig voneinander rekonstruiert und betrachtet. Problematisch sind auch die sehr unterschiedlichen Zeitskalen der einzelnen Netzwerktypen. Der Metabolismus passt sich wesentlich schneller an, als dies die Regulation tut. Dies erschwert die Integration der Netzwerktypen zusätzlich.

Die Computermodelle, die erstellt werden, können in zwei Typen unterschieden werden: Chemisch-kinetische Modelle, die durch ein System von chemischen Gleichungen erstellt werden; und diskrete Schaltkreismodelle, die ein Netzwerk von Knoten und gerichteten Beziehungen zwischen diesen Knoten darstellen.

1.5 Genomannotation

Um ein metabolisches Netzwerk zu rekonstruieren, ist es nötig zu wissen, welche Reaktionen in einem Organismus katalysiert werden können. Einen Teil dieser Information bietet der Enzymumfang des Organismus. Von diesem nicht erfasst werden Transportreaktionen und spontane Reaktionen.

Die Information, welche Enzyme von einem Organismus gebildet werden können, liegt auf dem Genom. Diese Informationen lassen sich in Datenbanken finden, die Genomannotationen anbieten. Unter dem Begriff Genomannotation fasst man alle zusätzlichen Informationen zusammen, die den reinen Sequenzen zugewiesen werden. Für die Rekonstruktion des Stoffwechselnetzwerkes sind die Enzymannotationen von Bedeutung.

Es gibt eine Vielzahl von Datenbanken, die diese Information für ein breites Spektrum von Organismen anbieten. Meist genutzt sind hier die RefSeq Datenbank des National Center for Biotechnology Information (NCBI)[4], die Kyoto Encyclopedia of Genes and Genomes (KEGG)[5, 6], die Proteindatenbank PEDANT[7] und die UniProtKB Datenbank[8].

1.5.1 Genomannotation in Datenbanken

Da es weitreichende Informationen oft nur zu Modellorganismen gibt, basiert nur ein kleiner Teil der Annotationen dieser Datenbanken auf manuell überprüften Daten. Damit auch die Genome anderer Organismen untersucht werden können, nutzen die Datenbanken Informationen aus Programmen, die Annotationen durch Computerberechnungen erstellen.

Auch wenn diese computerbasiert berechneten Daten von großem Wert sind, sind sie mit dem Problem behaftet, dass die Qualitätskriterien, mit denen sie erstellt wurden, oft nicht bekannt sind. Dies macht ihre Einordnung und Bewertung schwierig. Ohne weitreichende manuelle Auswertung ist es unmöglich, die wahren Gründe für bestehende Unterschiede zu ermitteln.

Allerdings existieren erhebliche Unterschiede zwischen Datenbanken, beispielsweise Stoffwechseldatenbanken[9]. Gleiches gilt für Datenbanken, die Informationen zu Enzymfunktionen liefern[10].

Teilweise lassen sich diese Unterschiede dadurch erklären, dass die automatische Zuweisung von Enzymfunktionen nach wie vor eine große Herausforderung darstellt[11]. Eine weitere Quelle der Unterschiede liegt in den unterschiedlichen Zeitpunkten, zu denen die Annotationen teilweise erstellt wurden. Somit beruhen sie auf verschiedenen Wissensständen.

Neue Zuweisungen von Enzymfunktionen beruhen oft auf fragwürdigen früheren Zuweisungen, die durch Fehler in der manuellen Annotation oder durch falschen Transfer von Informationen entstanden sind[10]. Dies alles führt zu erheblichen Unstimmigkeiten in der Genomannotation eines Organismus.

1.5.2 Genomannotation durch Computerprogramme

Neben den erwähnten Datenbanken, die Informationen zum Enzymbestand eines Organismus liefern, gibt es auch Programme, die dies tun. Gliedern kann man diese Programme in drei Gruppen: Solche, die eine Genomannotation erstellen, solche, die bestehenden Annotationen einen Qualitätswert zuweisen, und solche, die verschiedene Annotationsquellen integrieren.

Zur ersten Gruppe gehört PRIAM[12] Dieses Programm sagt Enzymfunktionen basierend auf einer Sammlung von Profilen vorher, die mit den Einträgen der ENZYME[13] Datenbank erstellt wurden. EFICAZ[14, 15] gehört ebenfalls zur ersten Gruppe und basiert ebenfalls auf Mustern. Es kann als Stand-alone-Programm heruntergeladen werden oder mittels eines Webinterfaces genutzt werden. EFICAZ bietet die Möglichkeit, Annotationsdaten aus einer externen Quelle zu integrieren. Allerdings kann zur Zeit nur KEGG eingebunden werden und keine anderen Quellen.

Zur zweiten Gruppe gehört der Ansatz, den Yang et al. vorschlagen[16]. Bei diesem Ansatz wird existierenden Annotationen ein Qualitätswert basierend auf dem Vergleich mit einigen Referenzorganismen zugewiesen. Ebenfalls zur zweiten Gruppe gehört das

Programm, das von Chitale et al.[17] entwickelt wurde. Es wird eine Annotation mit einem dazugehörigen Qualitätswert geliefert. Allerdings muss die Analyse Sequenz für Sequenz durchgeführt werden, was sich als sehr aufwendig herausstellen kann.

Zur dritten Gruppe gehören Apollo[18] und die UCSC Genome Browser Datenbank[19]. Mit beiden Programmen ist es möglich, Annotationsquellen zu integrieren. Allerdings werden nur Daten zu Genpositionen gesammelt und verglichen und nicht die verfügbaren Funktionsvorhersagen.

Somit haben Wissenschaftler heute meist nur die Wahl zwischen zwei Vorgehensweisen: Entweder sie entscheiden sich für eine Datenbank und akzeptieren damit, dass bestimmte Informationen verloren gehen, oder sie vergleichen die existierenden Annotationen per Hand.

Problematisch ist außerdem, dass der Aktualisierungszyklus der organismenspezifischen Genomannotationen trotz regelmäßiger Aktualisierung oft zu lang ist. Dies ist aber auf Grund der großen Datenmengen oft nicht anders von den Datenbanken zu realisieren.

1.6 Probleme biologischer Datenbanken

Bei der Erstellung von metabolischen Modellen hängt die Genauigkeit dieser Modelle zu großen Teilen von der Qualität der primären Ressourcen und der Genvorhersage ab[20]. Schon eine fehlende Enzymfunktion kann sich als kritisch erweisen und einen großen Einfluss auf das ganze Modell haben. Schnoes et al. haben festgestellt, dass Annotationsfehler in öffentlich zugänglichen Datenbanken ein nicht zu unterschätzendes Problem sind, da diese sich über die Zeit fortpflanzen können[21]. Doch auch, wenn Wissenschaftler versuchen, dies zu umgehen, indem sie mehrere Quellen hinzuziehen, bleibt das Problem bestehen, dass bei Unterschieden entweder eine aufwendige manuelle Untersuchung durchgeführt oder zufällig eine der bestehenden Informationen bevorzugt werden muss. Um dieses Problem zu beheben und eine guten Überblick über die bestehende Information zu ermöglichen, ist es erforderlich, spezialisierte Programme zu entwickeln, die annotieren und integrieren[20]. Aus diesem Grund wurde im Rahmen dieser Arbeit das Programm EnzymeDetector entwickelt.

1.7 EnzymeDetector

EnzymeDetector bietet ein rationales und umfassendes Bewertungsschema und sammelt die Information aller großen Annotationsdatenbanken, einer regelmäßig aktualisierten BLAST-basierten Genomannotation und einer Sequenzmustersuche. Somit liefert es einen schnellen Überblick über die möglichen Annotationskandidaten für jedes Gen und bietet darüber hinaus ein Maß für die Qualität der Annotationen.

Der Vorteil gegenüber den Datenbanken und Programmen, die Genomannotationen bieten, liegt darin, dass nicht jede Sequenz einzeln untersucht werden muss, sondern die Daten ganzer Organismen vorberechnet in einer Datenbank zur Verfügung gestellt werden. Dadurch entstehen keine Rechenzeiten zum Abrufen der Daten. Des Weiteren wurde ein einfacher Zugang zu dieser Datenbank geschaffen, und alle Informationen können vom Nutzer lokal gespeichert werden.

Bis jetzt wurden nur Organismen mit kleinen Genomen, genauer Prokaryoten untersucht. In Zukunft sollen die Ergebnisse um Annotationen von Eukaryoten erweitert werden.

Das Programm EnzymeDetector kann auch von Wissenschaftlern genutzt werden, die sich selbst nicht mit Bioinformatik beschäftigen, da der Zugang zu den Daten einfach über ein Webinterface realisiert worden ist.

1.8 Lückensuche

Selbst mit den integrierten Daten von Programmen wie dem EnzymeDetector verbleiben oft noch Lücken im Stoffwechselnetzwerk, die geschlossen werden müssen. Solange kritische Lücken nicht geschlossen werden, kann kein funktionierendes und ausgeglichenes Modell erstellt werden. Um diese Lücken zu schließen, können Programme wie der gapFiller[22] genutzt werden. Dieses Programm führt eine Analyse des mit Hilfe der EnzymeDetector-Informationen rekonstruierten Stoffwechselnetzwerkes durch. Es wird dabei getestet, ob alle Biomassekomponenten erreichbar sind. Das heißt, es muss gewährleistet sein, dass vom Ausgangsstoff, beispielsweise aus Glucose, mit Hilfe der vorliegenden Enzymfunktionen alle

Komponenten der Biomasse hergestellt werden können. Hierfür wird eine Kürzester-Pfad-Suche durchgeführt. Ist eine Komponente nicht zugänglich, wird überprüft, welche Enzymfunktionen für die Vervollständigung des betreffenden Stoffwechselweges fehlen. Anschließend wird gezielt nach dieser Enzymfunktion auf dem Genom gesucht.

1.9 Atomzuordnung

Um die Ergebnisse eines Programmes wie des gapFillers zu verbessern und zu verhindern, dass biologisch unsinnige Pfade beschritten werden, gibt es die Möglichkeit, die erlaubten Pfade mit Hilfe von Atomzuordnungsprogrammen einzuschränken. Für die Zuordnung der Atome der Reaktanden auf die Atome der Produkte bieten sich zwei Hauptansätze, graphbasiert oder bedingungsbasiert[23].

Der bedingungsbasierte Ansatz fordert stöchiometrische Ausgeglichenheit und hat den Nachteil, dass das resultierende Problem NP-vollständig ist. Somit ist eine Realisierung auf Genomgröße nicht möglich.

Beim graphbasierten Ansatz wird ein linearer Biotransformationsweg erstellt. Die Gültigkeit eines Weges wird nur darüber bestimmt, ob ein Atom aus dem Ausgangsstoff im Produkt vorhanden ist. Somit ergeben sich leider auch biologisch unsinnige Wege. Auch werden meist zu viele Wege gefunden. Beim graphbasierten Ansatz handelt es sich in den meisten Fällen um einen „Maximal common subgraph“-Algorithmus (MCS). Für diesen werden zwei Moleküle in Graphen umgewandelt, deren größte gemeinsame Teilstruktur ermittelt wird. Da auch dieses Problem NP-vollständig ist, werden meist heuristische Algorithmen eingesetzt[24].

Da die meisten MCS-Algorithmen eher zu viele Wege finden, wurde von Blum und Kohlbacher diese Art von Algorithmus nicht mit einer Kürzester-Weg-Suche kombiniert, sondern mit einer „lightest path“-Suche[23]. Dadurch wird verhindert, dass über Pool-Metabolite wie ATP Wege ermöglicht werden, die biologisch nicht sinnvoll sind. Es wird der Pfad gesucht, der die geringsten Kosten hat. Jeder Knoten erhält einen Kostenwert entsprechend den Kanten, die mit ihm verbunden sind. Doch auch dieses Vorgehen hat Nachteile, nämlich wenn diese Pool-Metabolite als Intermediate in Reaktionen vorkommen.

Die Server SIMCOMP und SUBCOMP[25] bieten die Möglichkeit, entweder über eine chemische Ähnlichkeitssuche graphbasiert nach den größten Cliquen zweier Moleküle zu suchen bzw. über einen Bitvektorvergleich eine chemische Substruktursuche durchzuführen. Eine Clique ist eine Menge der Knoten in einem ungerichteten Graphen, die alle paarweise miteinander verbunden sind. Für die Untersuchung zweier Moleküle wird ein Produktgraph aus beiden Molekülgraphen erstellt, in dem alle möglichen Überschneidungen der Knoten der Einzelgraphen in einem neuen Knoten des Produktgraphen repräsentiert werden. Bei einem Bitvektor wird das Molekül in einem Array dargestellt. Das Array besteht aus einer Reihe von Elementen. Diese Elemente können nur zwei Werte annehmen, eins und null. Nach Übertragung der Moleküle in ihre jeweiligen Bitvektoren können diese miteinander verglichen werden. Hierbei ist der graphbasierte Ansatz zwar akkurater als der bitstringbasierte, allerdings ist er auch wesentlich zeitaufwendiger. Nachteil beider Server ist, dass wie bei vielen anderen vorhandenen Algorithmen nur Einzelmoleküle verglichen werden und nicht alle Moleküle einer Reaktion.

Auf Grund dessen wurde für den gapFiller im Rahmen dieser Arbeit das bestehende Programm AtomMapper[26] erweitert. Dieses bietet die Möglichkeit über zwei kombinierte MCS-Algorithmen nach der größten Übereinstimmung zwischen Molekülen zu suchen. Die Ergebnisse dieser Suchen werden in einer Tiefensuche so kombiniert, dass die optimale Paarung für eine Reaktion gefunden wird.

1.10 Flux-Balance-Analyse

Durch die Kombination der vorgestellten Programme wird die Rekonstruktion von metabolischen Netzen enorm erleichtert. Die erstellten Modelle können genutzt werden, um eine Flux-Balance-Analyse (FBA) durchzuführen. Heutzutage werden mit dieser Methode „whole genome“-Netzwerke aus allen bekannten Reaktionen eines Organismus analysiert[27]. Allerdings können auch isolierte Teile, wie es bei Einführung der Methode üblich war, untersucht werden. Bei der FBA handelt es sich um einen mathematischen Ansatz zur Vorhersage der Flüsse durch ein metabolisches Netzwerk. Diese Analysemethode kann u.a. genutzt werden, um die Wachstumsrate eines Organismus vorherzusagen oder die Produktionsrate eines interessanten

Metaboliten. Im ersten Schritt der FBA wird die stöchiometrische Matrix des Netzwerks aufgestellt. Jede Zeile dieser Matrix entspricht einem Metaboliten und jede Spalte einer Reaktion. Die Elemente der Matrix entsprechen den stöchiometrischen Koeffizienten der Metabolite in der Reaktion. Für die FBA wird unter Annahme des Fließgleichgewichtes eine Zielfunktion definiert, die minimiert oder maximiert werden soll. Eine Zelle bzw. ein Organismus befindet sich im Fließgleichgewicht, wenn alle Metabolite in gleichem Maße auf-, wie abgebaut werden. Somit ändert sich die absolute Konzentration der Metabolite über die Zeit nicht. Gelöst wird dieses mathematische Optimierungsproblem mit Hilfe der Linearen Programmierung. Zusätzlich zur Bedingung des Fließgleichgewichts wird der Lösungsraum häufig durch obere und untere Grenzen für manche Reaktionsflüsse eingeschränkt.

Neben der Vorhersage von Wachstums- und Produktionsraten kann mit Hilfe der FBA und verwandten Methoden, wie der „Minimization of Metabolic Adjustment“ (MOMA), das Verhalten von Knockout-Mutanten vorhergesagt werden.

So können teure und zeitaufwendige Laboranalysen vermieden werden. Des weiteren ist es unmöglich, alle metabolischen Flüsse eines Organismus im Labor zu bestimmen.

1.11 Visualisierung von Stoffwechselnetzwerken

Zur Veranschaulichung der Ergebnisse der verschiedenen Analyse- bzw. Modellierungsschritte, also der reinen EnzymeDetector-Ergebnisse, der Ergebnisse der kombinierten EnzymeDetector/gapFiller-Analyse und der rekonstruierten metabolischen Netzwerke bzw. der Ergebnisse der darauf durchgeführten FBA, ist es sinnvoll, die Daten zu visualisieren. Dies vereinfacht die Analyse und erleichtert das Auffinden von Lücken oder Fehlern. Auch ist das Nachvollziehen von Stoffwechselflüssen in einer grafischen Darstellung wesentlich übersichtlicher und einfacher, als wenn dies nur anhand von Reaktionslisten und zugehörigen Flusszahlen geschehen müsste.

Den gesamten Stoffwechsel eines Organismus als zusammenhängende Karte gibt es so aber nicht. KEGG[5] bietet zwar eine Gesamtübersicht an, die Stoffwechselwege im Genaueren sind aber auf einzelne Karten aufgeteilt. Dies erschwert die Übersicht. Auch sind die Stoffwechselwege auf diesen Karten oft nicht so angeordnet, wie man es aus

Lehrbüchern gewohnt ist, was das Zurechtfinden deutlich erschwert. Ein weiterer Nachteil ist, dass es zwar möglich ist, die Karten nach Organismen einfärben zu lassen, allerdings liegen diesen organismenspezifischen Karten nur die KEGG-Informationen zu Grunde. Dadurch gehen viele Informationen verloren. Informationen, die z.B. nur aus der Primärliteratur zu ermitteln sind, können so nicht in die Analyse eingebunden werden. Ein weiterer Nachteil der KEGG-Karten ist, dass kaum hinterlegt ist, aus welchen Quellen die Informationen zu den Stoffwechselreaktionen stammen. So lässt sich nicht nachvollziehen, welche Güte die eingetragenen Reaktionen haben.

Anders ist dies in MetaCyc[28]. Hier finden sich zu allen eingetragenen Reaktionen Literaturverweise. Nachteil ist jedoch auch hier, dass es sich um einzelne Teilkarten handelt und keinen zusammenhängenden Graphen. Die Informationen lassen sich zwar organismenspezifisch anzeigen, doch auch hier werden nur die internen Stoffwechselwege und Annotationen zugelassen und es können keine weiteren Informationen hinzugefügt werden.

Aus diesem Grund wurde im Rahmen der Arbeit an der Planung und Erstellung einer generischen Stoffwechselkarte mitgearbeitet. Bei dieser handelt es sich um einen zusammenhängenden Graphen. Die Ergebnisse aller Programme, die in dieser Arbeit vorgestellt werden, können auf dieser Karte visualisiert werden. Beim manuellen Erstellen der Karte wurde Wert darauf gelegt, eine Anordnung zu erhalten, die ähnlich der Anordnung der Stoffwechselwege in Lehrbüchern ist.

Der Zusammenschluss aller Programme und Tools erleichtert somit das Vorgehen von der reinen Genomsequenz zum vollständigen Modell eines Organismus, welches analysiert und visualisiert werden kann.

2 Daten, Algorithmen und Methoden

2.1 EC-Nummern

Die „Enzyme Commission (EC)“-Nummern bilden eine numerische Klassifikation von Enzymen nach den durch sie katalysierten biochemischen Reaktionen. Sie bestehen aus vier Zahlen, die durch Punkte voneinander getrennt sind. Die erste Ziffer teilt das Enzym in eine der sechs Enzymhauptklassen ein. In Tabelle 1 sind die Hauptklassen und deren Kurzbeschreibungen aufgelistet.

Tabelle 1 – Auflistung der sechs Enzymhauptklassen, ihre Bezeichnung und die Art der Reaktionen, die durch sie katalysiert werden

Hauptklasse	Bezeichnung	katalysierte Reaktionsart
1	Oxidoreduktasen	katalysieren Redoxreaktionen
2	Transferasen	katalysieren die Übertragung von funktionellen Gruppen
3	Hydrolasen	katalysieren die hydrolytische Spaltung
4	Lyasen	katalysieren die nicht-hydrolytische Eliminierung oder Addition von Atomgruppen
5	Isomerasen	katalysieren die intramolekulare Umwandlung (Isomerisierung)
6	Ligasen	katalysieren die Bildung kovalenter Bindungen mit Hilfe energiereicher Cofaktoren

Die weiteren Zahlen kategorisieren Informationen zur Reaktion wie Angriffspunkt des Enzyms (welche Art von Bindung), Nebenprodukt/e der Reaktion, erforderliche Reaktionspartner , Cofaktoren usw. Ist eine Reaktion noch nicht zur Gänze kategorisiert, wird eine unvollständige EC-Nummer zugewiesen, bei der nicht alle 4 Zahlen eindeutig bestimmt sind. Sind alle 4 Zahlen zugewiesen, handelt es sich um eine eindeutige Identifizierung eines Reaktionsmechanismus[29].

Die enzymatische Aktivität von Proteinen ist in den in der Arbeit genutzten Datenbanken in Form der EC-Nummern hinterlegt. Dies ermöglicht einen eindeutigen Vergleich der Informationen. Wären nur die Namen der Enzyme hinterlegt, wäre der Abgleich der Informationen um einiges schwieriger, da meist mehrere Namen für ein Enzym bestehen.

2.2 Operon

Bei einem Operon handelt es sich um eine funktionelle Einheit des Genoms von Prokaryonten. Es besteht aus mehreren Strukturgenen, einem Operator und einem Promotor. Der Promotor ist die Bindungsstelle für die RNA-Polymerase. Der Operator bietet die Möglichkeit zur Regulation des gesamten Abschnitts, da an ihn Induktoren oder Repressormoleküle binden können. Dadurch wird die Bindung der RNA-Polymerase erleichtert oder behindert. Meist stehen Gene, die im selben Operon liegen, in einem funktionellen Zusammenhang, wie zum Beispiel einer Stressantwort[30].

Diese Eigenschaft nutzt die im Rahmen des EnzymeDetectors durchgeführte Operonanalyse. Diese wird genutzt, um bestehende Annotationen mit Hilfe der Operoninformationen bzw. dem funktionellen Zusammenhang der Gene zu bewerten und besser einordnen zu können.

2.3 BLAST-Algorithmus

Im Rahmen des Programms EnzymeDetector wird das Basic Local Alignment Search Tool (BLAST) des NCBI ausgeführt (Version 2.2.24)[31]. Das Programm kann entweder auf der Website ausführt werden oder (wie in der Arbeit genutzt) als Stand-alone-Version heruntergeladen werden.

Der BLAST-Algorithmus führt ein lokales Alignment, d.h. eine Ähnlichkeitssuche einer Query-Sequenz gegen eine Datenbank, durch. Bei einem Alignment handelt es sich um das Ergebnis eines Verfahrens zum paarweisen Sequenzvergleich, das darauf abzielt, in beiden Sequenzen vorkommende Module zu identifizieren und deren Position zu bestimmen.[32]

Bei BLAST handelt sich um ein heuristisches Programm, das als Ergebnis zwei statistische Werte liefert. Der eine Wert sagt aus, wie ähnlich sich zwei Sequenzen sind (genannt Bit Score). Der andere Wert gibt an, mit welcher Wahrscheinlichkeit eine Sequenz mit einem mindestens so hohen Bit Score wie dem vorliegenden durch Zufall in einer Datenbank einer bestimmten Größe gefunden wird (genannt Expect (E)-value).

Je kleiner der E-value ist, desto besser ist der zugehörige BLAST-Treffer. Auf Grund dessen kann dieser Wert auch sehr gut als Signifikanzschwellenwert genutzt werden.

Der Algorithmus geht den Sequenzvergleich in 3 Schritten an:

1. Entsprechend der gewählten Wortlänge x (für die Arbeit wurde der Standardwert von 3 beibehalten) wird eine Tabelle mit allen möglichen Wörtern der Länge x aus der Query-Sequenz hergestellt.

2. Die Suchdatenbank wird nach Vorkommen der ermittelten Wörter durchsucht.

3. Wurde ein Wort gefunden, wird das Alignment maximal ausgedehnt.

Es gibt verschiedene BLAST-Unterprogramme, die genutzt werden können, abhängig davon, welche Art von Sequenz vorliegt und gegen welche Art von Sequenzdatenbank gesucht werden soll. Man kann entweder mit Nukleotidsequenzen oder mit Proteinsequenzen suchen, und diese beiden Möglichkeiten bieten sich auch für die Datenbank. In der vorliegenden Arbeit wurde das Unterprogramm blastp genutzt. Dieses Programm ermöglicht die BLAST-Suche von Proteinsequenzen gegen eine Proteindatenbank. Dies bot sich an, da in den Genomannotationen der Datenbanken die „gene identifier“ der Proteinsequenzen vorliegen. Um die Integration aller Daten zu gewährleisten, muss mit diesen Sequenzen geblastet werden. Da gegen die UniProt Datenbank gesucht werden sollte, bei der es sich um eine Proteindatenbank handelt, war das blastp Programm die logische Wahl.

Für die BLAST-Suche ist es erforderlich, die Sequenzdatenbank in einem bestimmten Format vorliegen zu haben. Die UniProt-Datenbank ist allerdings nur im FASTA-Format verfügbar. Somit muss die Datenbank vor dem Programmstart mit Hilfe des Programms formatdb in das erforderliche Format umgewandelt werden.

Die vom EnzymeDetector durchgeführte BLAST-Suche stellt den zeit-limitierenden Faktor des Gesamtprogramms dar. Pro Genabschnitt braucht eine BLAST-Suche, wenn nur ein Prozessor genutzt wird, ca. 80 Sekunden. Um das Programm zu optimieren, wird, wenn möglich, mit 4 Prozessoren gerechnet. Dies verringert die Dauer einer BLAST-Suche auf 19 Sekunden. Da aber prokaryotische Organismen, die bis jetzt

ausschließlich mit EnzymeDetector untersucht wurden, im Schnitt ca. 2000 Gene haben, dauert die Durchführung einer BLAST-Suche für das gesamte Genom durchschnittlich immer noch ca. 10 Stunden. Für ein vollständiges Update der Programmergebnisse müssen ca. 1500 prokaryotische Organismen neu berechnet werden. Um den Zeitaufwand dieses EnzymeDetector-Updates trotzdem in einem vernünftigen Rahmen zu halten, wird das Programm parallelisiert auf einem Rechencluster ausgeführt, wie unter Punkt 2.11 beschrieben.

2.4 BrEPS

Beim Braunschweig Enzyme Pattern Search (BrEPS) handelt es sich um ein Programm, das eine Genomannotation mit Hilfe einer Mustersuche durchführt[33]. Es besteht aus drei Abschnitten, die nacheinander ausgeführt werden.

In einem ersten Abschnitt werden alle Einträge des Swiss-Prot Anteils der UniProtKB in drei Kategorien eingeteilt und nur die Kandidaten aus der Gruppe Enzyme werden für die Berechnungen verwendet. Diese strikte Auswahl, bei der keine putativen Enzymfunktionen zugelassen werden, führt dazu, dass am Ende der Berechnungen hoch-qualitative Muster entstehen. Die gefilterten Sequenzen werden in einem All-vs.-all-BLAST miteinander verglichen.

Die erhaltenen E-values werden in einem zweiten Abschnitt als Distanzmaß für ein Complete Linkage Clustering genutzt. Hierbei handelt es sich um einen iterativ vorgehenden Algorithmus, der Objekte in Clustern ordnet. Es werden immer die Objekte zusammengefügt, die die geringste Distanz zueinander aufweisen. Für das neu entstandene Objekt werden wieder die Distanzen zu den anderen Objekten ermittelt und die Objekte geringster Distanz zusammengeführt. Die entstehenden Musterknoten werden als Bäume gespeichert.

Im letzten Abschnitt des Programms werden aus den gesammelten Sequenzen der entstandenen Baumknoten mit Hilfe von ClustalW Muster erstellt.

Das Programm liefert als Ergebnis Muster verschiedener Güte mit einer unterschiedlichen Anzahl von EC-Nummern, die durch das Muster abgedeckt werden.

Die Güte der Muster wird beim Durchlaufen des Programms berechnet und beim Ergebnis hinterlegt. Gütekriterium für diese Auswertung ist hierbei die Übereinstimmung mit BRENDA.

Die BrEPS-Ergebnisse werden für die einzelnen Organismen in die Ergebnisse des Programms EnzymeDetector integriert. Hierbei werden den Ergebnissen abgestufte Relevanzwerte je nach Gütewert aus BrEPS zugewiesen.

2.5 Datenbanksystem

Auf das Datenbanksystem wird per Structured Query Language (SQL) zugegriffen. SQL ist eine Datenbanksprache, mit der Daten in einem relationalen Datenbankmodell definiert, abgefragt und gesichert werden können. Grundsatz ist es, dass jede Information nur einmal in der Datenbank erfasst wird und dass mit Hilfe von eindeutigen Schlüsseln auf diese Inhalte verwiesen werden kann. Neben der einfach zu lernenden Syntax ist ein großer Vorteil von SQL, dass viele Programmiersprachen und Anwendungen vordefinierte Schnittstellen für den Zugriff per SQL anbieten.[34]

2.5.1 EnzymeDetector-Programm

Für die Speicherung der Ergebnisse und das Sichern von Zwischenergebnissen des Programms EnzymeDetector wurden SQL-Datenbanken angelegt.

Für die Daten, die zur Berechnung der Ergebnisse erforderlich waren, wurde die SQL-Datenbank „Data_Input_for_Modelling“ angelegt. Diese Datenbank enthält die in Tabelle 2 aufgeführten Tabellen. Diese Tabellen werden in regelmäßigen Abständen (alle 6 Monate) im Zuge eines Updates und jedes Mal, bevor die Berechnungen des EnzymeDetectors durchgeführt werden, aktualisiert. So ist gewährleistet, dass die Berechnungen stets auf Daten basieren, die den neuesten Stand haben.

Die Endergebnisse des Programms werden in der SQL-Datenbank „Metabolic_Models“ abgelegt. Tabelle 3 gibt eine Übersicht über die umfassenden Tabellen und Informationen.

Tabelle 2 - Auflistung der SQL-Tabellen der Datenbank „Data_Input_for_Modelling“, die zur Berechnung des Programms EnzymeDetector benötigt werden, ihr Inhalt und die Anzahl der Einträge in diesen Tabellen

Tabellenname	**Tabelleninhalt**	**Anzahl Einträge**
EC_Info	Liste aller EC-Nummern aus BRENDA mit Informationen dazu, ob die EC-Nummer eventuell gewechselt hat oder komplett gelöscht wurde	5411
KEGG_Organism	Alle in KEGG vorkommenden Organismen gespeichert mit Namen, „taxonomy ID“, Dreibuchstaben-Code, NC-Nummer, „Refseq-ID“ und „T-Nummer“	1536
NCBI_Sequenzen	Alle Proteinsequenzen, die in NCBI hinterlegt sind und deren GI	31445836
Uniprot_Enzyme	Link zwischen den UniProtKB-Accession-Nummern und den EC-Nummern	1992080
UniProt_to_NcbiGI	Link zwischen den Accession-Nummern aus der UniProtKB und der GI aus NCBI	27747924

Tabelle 3 - Auflistung der Tabellen der Datenbank „Metabolic_Models“ und der darin enthaltenen Informationen. Diese Tabellen werden für Berechnungen des Programms EnzymeDetector benötigt und speichern die Endergebnisse

Tabellenname	**Tabelleninhalt**
BREPS_Relevanzen	Relevanzwerte der einzelnen BrEPS-Gruppen (abgestuft nach Gütewert)
Gene	Sammlung aller Gene aller Organismen, die bisher berechnet wurden, mit Informationen zu GI, Locus Tag, Organismus und Gen-Start und -Stop
Input_Databases	Sammlung der Annotationsquellen und ihrer Standardrelevanz
ownBLAST_Relevanzen	Standartrelevanzen der einzelnen Gütegruppen der BLAST-basierten Genomannotation und ihre Grenzwerte
Species	Liste aller Organismen der Domäne Bacteria und Archaea aus KEGG_Organisms mit den unterschiedlichen Flags aus dem Programm EnzymeDetector und der Bezeichnung der Ergebnistabelle. Die Flags umfassen die Information, ob eine BLAST-Suche durchgeführt wurde, ob die Daten freigegeben sind und ob der Organismus neu berechnet werden soll oder nicht.
BrEPS__ *Organismenbezeichnung*	Tabelle mit den BrEPS-Ergebnissen des Organismus
ED__ *Organismenbezeichnung*	Ergebnistabelle des EnzymeDetector

2.5.2 EnzymeDetector-Website

Für das Webinterface des Programms EnzymeDetector wurde eine Spiegelung der Tabellen, die unter Punkt 2.5.1 beschrieben wurden, erzeugt. Dies war nötig, da die Website auf einem anderen Server liegt. Zum anderen ist auf diese Weise eine strikte Trennung zwischen Backend und Webserver gewährleistet. Es ist somit schon auf physischer Ebene ausgeschlossen, dass über die Webseite auf Daten zugegriffen wird, die gerade in der Berechnung sind oder erst noch geprüft werden müssen.

2.5.3 Stoffwechselkarte

Die im Rahmen dieser Arbeit mit geplante und erstellte Stoffwechselkarte wurde zwar mit Hilfe eines Netzwerk-Visualisierungsprogramms als XGMML-Datei (siehe Punkt 2.20 auf Seite 44) erstellt, allerdings wurde ein Programm geschrieben, das die Stoffwechselkarte in eine SQL-Datenbank überträgt. Die Bearbeitung und Abfrage der Daten wird erleichtert, da dies so auch auf Datenbankebene möglich ist. Die Planung und Erstellung der Datenbank fand im Rahmen eines studentischen Etagenpraktikums statt, welches in dieser Arbeit betreut wurde. In Abbildung 1 sind die wichtigsten Tabellen der Datenbank „metabolic_pathways“ aufgeführt und ihre Beziehungen untereinander veranschaulicht.

Jeder Knoten des Netzwerkes ist durch einen Eintrag in der Tabelle „node“ und jede Kanten in der Tabelle „edge“ repräsentiert. In diesen Tabellen finden sich die grafischen Informationen zu den Knoten-Objekten wie die x- und y-Position im Netzwerk, die Größe und Form des jeweiligen Knotens, die Beschriftung u.a.. Durch Verbindungstabellen sind den Knoten-Einträgen einerseits grundlegende Informationen zugeordnet, wie, um welche chemische Verbindung bzw. welches Enzym es sich handelt. In den entsprechenden Tabellen „substance“ und „enzyme“ finden sich Informationen zu den Verbindungen bzw. Enzymen selbst. Auf der anderen Seite finden sich auch Verknüpfungen zu weiterführenden Informationen wie die Zuordnung zu einer Reaktion oder einem Stoffwechselweg und die Zuordnungen zu den Referenzen aus den Stoffwechselweg-Quellen.

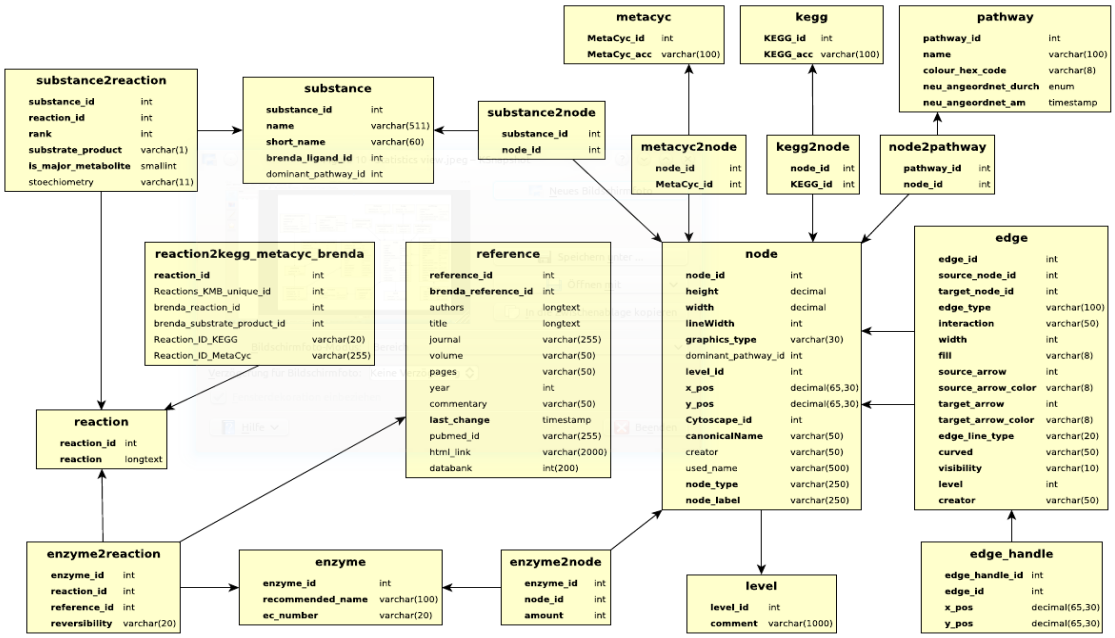


Abbildung 1 - Tabellen der SQL-Datenbank „metabolic_pathways“ und ihre Beziehungen untereinander (Abbildung aus [57] übernommen) (Abbildung groß unter Anhang 1)

Aus der Gesamtheit der Tabellen lässt sich mit Hilfe eines dafür geschriebenen Programms wieder eine XGMML-Datei erstellen, welche das Netzwerk-Visualisierungsprogramm darstellen kann.

2.6 Hintergrund-Datenbanken für den EnzymeDetector

Für die Durchführung des Programms EnzymeDetector ist es erforderlich, Informationen aus verschiedenen biologischen Datenbanken auszulesen und zu speichern. Diese Informationen müssen in regelmäßigen Abständen im Zuge eines Updates erneut abgefragt werden (jedes halbe Jahr). Dafür wurde ein automatisches Skript entwickelt, das die Informationen aus den Datenbanken abfragt und in SQL-Tabellen ablegt. Eine solche Aktualisierung erfordert ca. zwei Tage Rechenzeit auf Grund der Datenmengen, die hier heruntergeladen und prozessiert werden müssen. Im Folgenden sind die Datenbanken aufgeführt, die im Rahmen eines solchen Updates ausgelesen werden.

2.6.1 KEGG Organism

Die Kyoto Encyclopedia of Genes and Genomes (KEGG)[5, 6] ist eine integrierte Datenbank, die aus 16 Hauptdatenbanken besteht. Sie bietet eine Vielzahl von genomischen und chemischen Informationen, ihr Hauptaugenmerk liegt aber auf der Bereitstellung von Stoffwechselweg-Karten. Aktuell (Stand Mai 2011, [35]) sind Informationen zu 1588 verschiedenen Organismen hinterlegt. Für die vorliegende Arbeit wird mittels des angebotenen FTP-Servers[1][36] die Taxonomie-Datei mit einer Liste all dieser Organismen abgefragt. So können jeweils die Namen, Abkürzungen, KEGG-Bezeichner und Links zu anderen großen Datenbanken gespeichert werden.

2.6.2 NCBI nr

Das National Center for Biotechnology Information (NCBI) ist Teil des National Institutes of Health (NIH) der USA und stellt eine Vielzahl von Datenbanken mit biologischen Informationen zur Verfügung[37]. Für die vorliegende Arbeit wurde eine nicht-redundante Sammlung aller Proteinsequenzen, die nr-Datenbank, ausgelesen und gespeichert. Diese Informationen werden benötigt, da im Rahmen des Programms EnzymeDetector eine BLAST-Suche (siehe Punkt 2.3) durchführt wird. Die aktuelle nr-Datenbank ist im gezipten Zustand unter [38] verfügbar.

2.6.3 UniProtKB

Die Universal Protein Resource (UniProt) ist eine Sammlung von Sequenzen mit zugehörigen Annotationen des Locus. Sie besteht aus der UniProt Knowledgebase (UniProtKB), den UniProt Reference Clusters (UniRef) und dem UniProt Archive (UniParc)[2]. Die UniProtKB kann unterteilt werden in die Swiss-Prot Datenbank und die TrEMBL-Datenbank, wobei Swiss-Prot nur manuell überprüfte Daten beinhaltet (aktuell 526.969 Sequenzeinträge, Stand Mai 2011), während die TrEMBL-Daten mit bioinformatischen Methoden automatisch generiert werden (aktuell 13.555.721 Sequenzen, Stand Mai 2011).

1 Zum Zeitpunkt des Einreichens dieser Arbeit waren die Daten, die KEGG anbietet, nicht mehr frei verfügbar. Das bedeutet, dass in allen Aktualisierungen die folgen diese Daten nicht mehr integriert werden können.

Die BLAST-Suche des Programms EnzymeDetector wird gegen die vollständige UniProtKB-Datenbank durchgeführt. Diese liefert daher als Ergebnis UniProt Accession-Nummern. Hinter diesen verbergen sich meist Enzymfunktionen. Da für die Modellierung EC-Nummern gebraucht werden, werden diese Accession-Nummern in EC-Nummern übersetzt. Die dafür erforderliche Verknüpfung zwischen EC-Nummern und Accession-Nummern wird aus der UniProtKB Datenbank ausgelesen und in einer SQL-Tabelle abgelegt.

Zusätzlich wird die vollständige UniProtKB gespeichert. Nach einer Umformatierung stellt diese Datenbank die Referenzdatenbank der durchgeführten BLAST-Suchen dar.

Die aktuelle UniProtKB kann mittels FTP-Server unter [39] abgefragt werden.

2.6.4 DOOR

In der Database of prOkaryotic OpeRons (DOOR)[40, 41] finden sich zur Zeit (Stand Mai 2011) die Operons von 971 verschiedenen Prokaryoten[42].

Erstellt werden diese Operons mittels eines bioinformatischen Vorhersage-Tools. Es handelt sich dabei um einen Data-Mining-Klassifikator. Der Vorhersage-Algorithmus arbeitet mit verschiedenen Datenarten, um eine hohe Genauigkeit zu gewährleisten. Es werden sechs verschiedene Ansätze bzw. Informationsquellen genutzt. Die intergenetische Distanz, die Nachbarschafts-Konservierung, die phylogenetische Distanz, das Vorkommen bestimmter kurzer DNA-Motive, das Ähnlichkeitsmaß der GO Terme der Gene und der Längenquotient zwischen den Genen werden beachtet.

Für die Nutzung der Informationen in der Operonanalyse des EnzymeDetectors werden die gesammelten Operons aus der Datenbank abgefragt. Dies geschieht über Speicherung eines TAR-Verzeichnisses, das unter [43] verfügbar ist. Dieses TAR-Verzeichnis beinhaltet einzelne Dateien mit den Operoninformationen der verschiedenen Organismen.

Im Rahmen der Operonanalyse wird die entsprechende Datei des untersuchten Organismus ausgelesen. Es werden ohne Einschränkungen alle Informationen aus der

Datei übernommen. Die ausgewerteten Informationen können dann in die Bewertung der Gesamtergebnisse des EnzymeDetector einbezogen werden, falls gewünscht.

2.7 Stoffwechselweg-Datenbanken

Für die Erstellung einer generischen Stoffwechselkarte (siehe Punkt 2.22 auf Seite 45) und die Bewertung der Operon-Informationen im Rahmen des Programms EnzymeDetector (siehe Punkt 2.13 auf Seite 36) wurden die Informationen der zwei großen Stoffwechselweg-Datenbanken KEGG und MetaCyc zur Orientierung zu Hilfe genommen.

2.7.1 KEGG

KEGG bietet eine Vielzahl von Informationen zu Metaboliten, Enzymen und Reaktionen. Hauptaugenmerk der Datenbank liegt aber auf den Stoffwechselwegkarten. Aktuell (Stand Juli 2011, [35]) bietet KEGG 400 Stoffwechselkarten an, die 143.793 Referenzen beinhalten. Diese Karten sind von Hand erstellt und bieten Informationen zum Metabolismus und zu molekularer Interaktion und Regulation. Es ist möglich, die Karten organismenspezifisch einfärben zu lassen.

Ein Nachteil der KEGG-Karten ist die etwas unübersichtliche Organisation und die willkürliche Aufteilung der Reaktionen auf die Karten, anders als aus Lehrbüchern und Fachpublikationen gewohnt. So fällt es oft nicht leicht, sich auf einer Karte zurechtzufinden.

2.7.2 MetaCyc

Bei MetaCyc[28] handelt es sich um eine Sammlung nicht-redundanter, experimentell aufgeklärter metabolischer Stoffwechselwege. Diese Datenbank enthält über 1.747 Stoffwechselwege aus mehr als 2.170 verschiedenen Organismen. Sie wird mittels wissenschaftlicher Literatur zu Experimenten manuell gepflegt. Die Wege, die in MetaCyc auf einer Ansicht zu sehen sind, sind meist viel kleiner als die KEGG-Karten. Dies steigert zwar die Übersichtlichkeit der Reaktionen, erschwert aber den Überblick über den Gesamtzusammenhang.

2.8 Vom EnzymeDetector genutzte Annotationsdatenbanken

Abgesehen von den Daten, die allgemein für die Berechnungen des EnzymeDetector benötigt werden, ist es außerdem erforderlich, für jeden Organismus die Genomannotationsdateien der verschiedenen Datenbanken herunterzuladen. Diese Dateien werden automatisch vor der Prozessierung des Organismus gespeichert, um immer die aktuellste Genomannotationsversion der einzelnen Datenbanken zum Zeitpunkt des Programmlaufs zur Verfügung zu haben.

Für die Arbeit wurden die größten Datenbanken genutzt, die diese Art von Information anbieten. Dies war erforderlich, da für die automatische Abfrage und Auswertung der Daten für jede Datenbank eine eigene Routine genutzt werden muss, die die Informationen herunterlädt und aus der Datei extrahiert. Somit wäre es sehr aufwendig, kleine Datenbanken, die nur wenige Organismen, teilweise sogar nur einen Organismus enthalten, in diesen Automatismus einzubinden, auch wenn die enthaltene Information sehr wertvoll ist. Die Daten wichtiger kleiner Datenbank können im Nachhinein per Hand eingebunden werden wenn dies gewünscht ist, zum Beispiel für Referenzorganismen bzw. Organismen mit speziellem Interesse für die Arbeitsgruppe.

2.8.1 NCBI RefSeq

Die Reference Sequence Datenbank (RefSeq) wird vom National Center for Biotechnology Information(NCBI) angeboten[4]. Sie wird ausschließlich aus Daten aus öffentlich zugänglichen Datenbanken erstellt. Das Hauptaugenmerk liegt darauf, eine akkurate, nicht-redundante und umfassende Sammlung natürlich vorkommender DNA-, RNA- und Protein-Moleküle zu schaffen. Auch wenn die biologischen Sequenzen aus GenBank-Einträgen stammen, handelt es sich bei den RefSeq-Einträgen um eine Synthese von Informationen. Es gibt vier Wege, auf denen Annotationen den reinen Sequenzinformationen hinzugefügt werden[37]:

1. Durch die Kollaboration mit weiteren Arbeitsgruppen, die selbst nicht am NCBI sitzen, aber eine manuelle Pflege der Daten vornehmen.

2. Durch eine automatische Genomannotations-Pipeline, die basierend auf Sequenzähnlichkeiten automatisch Annotationen berechnet.

3. Durch die Extraktion von GenBank-Einträgen, falls dort weiterführende Informationen von den Autoren eingetragen wurden.

4. Durch die manuelle Annotation durch NCBI Mitarbeiter.

Aktuell (Stand Mai 2011) sind Daten zu 6.190 verschiedenen mikrobiellen Taxonomie-IDs hinterlegt mit 9.591.571 Accession-Nummern zu Proteineinträgen[44].

Um die Genomannotationen verschiedener Organismen automatisiert herunterzuladen, wurde in der Arbeit der FTP-Server von NCBI genutzt. Unter [45] sind jeweils Ordner der verschiedenen Organismen hinterlegt. In diesen Ordnern findet sich die jeweilige Genomannotation, die gespeichert und ausgewertet wird.

2.8.2 KEGG GENES

Bei der Datenbank KEGG GENES[5] handelt es sich um einen Katalog kompletter Genome mit einer manuellen Genomannotation. Die Daten werden aus der RefSeq und anderen öffentlich zugänglichen Quellen generiert[46]. Zugänglich sind die Daten mittels eines FTP-Servers unter [47] und können so leicht automatisch aus dem Programm heraus abgefragt werden.

2.8.3 PEDANT

Die Datenbank des Protein Extraction, Description and ANalysis Tool (PEDANT)[7] bietet eine weitreichende automatische Analyse von genomischen Sequenzen durch eine Fülle von bioinformatischen Tools. Die Funktionsvorhersage im Speziellen wird durch eine hoch-stringente BLAST-Suche durchgeführt. Gesucht wird gegen Proteinsequenzen, die manuell in Funktionskategorien eingeordnet wurden, entsprechend dem Functional Catalogue, entwickelt vom Munich Information Center for Protein Sequences (MIPS) und der Biomax AG.[48]

Die Genomannotation des Organismus wird automatisch mittels eines SOAP-Clients abgefragt, der die gewünschten Daten zurückliefert. Der SOAP-Server ist unter [49] erreichbar. Dieser gibt die Abfrageergebnisse zurück und mit diesen kann dann im Programm weitergearbeitet werden.

2.8.4 UniProt ID mapping data

Um an die Annotationsinformationen der UniProtKB-Datenbank zu gelangen, wird die Datei „ID mapping data“ heruntergeladen und ausgelesen. Diese Datei enthält Links zwischen GIs und UniProt-Accession-Nummern. Aus diesen Accession-Nummern kann wiederum die Enzymfunktion abgeleitet werden, und man erhält die Information, welche Enzymfunktion welchem Genabschnitt in UniProtKB zugeordnet ist. Die Einträge in der „ID mapping data“-Datei werden alle vier Wochen aktualisiert und aus der UniProtKB extrahiert. Im Rahmen der Arbeit wurde nur der Teil der Informationen genutzt, der Swiss-Prot zugeordnet werden konnte. Nur diese Informationen sind manuell überprüft und von sehr hoher Qualität. Verfügbar ist die aktuellste Version der Datei unter [50].

2.9 Referenzorganismen

Die Daten der einzelnen Annotationsquellen, die in das EnzymeDetector-Ergebnis integriert werden, müssen in Relation zueinander gesetzt werden. Hierfür wurden neun ausgewählte Organismen (im weiteren Verlauf Referenzorganismen genannt) als Trainingsdatensatz bestimmt. Die Informationen der verschiedenen Quellen für diese neun Organismen wurde gegen ein Wahrheitsmaß ausgewertet und daraus wurden die Standard-Relevanzwerte für diese Quellen bestimmt.

Die neun Organismen sind in Tabelle 4 aufgelistet. Sie wurden ausgewählt, weil sie einen Querschnitt über die verschiedenen Prokaryotenarten bilden und unterschiedlich gut untersucht sind. Auch sind diese Organismen für unsere Arbeitsgruppe von großem Interesse. Im weiteren Verlauf werden nur noch die gängigen Kürzel für diese Organismen erwähnt, nicht mehr die vollständige Bezeichnung des jeweiligen Stamms. Gemeint sind aber immer die Organismen aus dieser Tabelle.

Bei der statistischen Auswertung des Programms EnzymeDetector werden die Ergebnisse nur für die Referenzorganismen explizit angegeben. Ansonsten werden Durchschnittswerte aufgeführt, die über alle 90 berechneten Organismen gemittelt wurden. Diese 90 Organismen gehören ausschließlich zu den Prokaryoten. Bisher wurden keine Eukaryoten-Annotationen erstellt. Dies ist der Fall, da einerseits für

unsere Arbeitsgruppe Prokaryoten eine größere Rolle spielen und andererseits die Berechnung der Annotationen unkomplizierter ist. Die Genome von Eukaryoten sind meist um ein Vielfaches größer. Hinzu kommt, dass sich für die Gene von Eukaryoten verschiedene Splicevarianten ergeben können, was die Anzahl der Sequenzen, die geblastet werden müssen, weiter erhöht.

Das untersuchte Genom der Organismen bildet den Querschnitt der entnommenen Informationen aus den integrierten Datenbanken. Es wurde keine eigenständige Zuweisung von Genloci durchgeführt, sondern die Sequenzen wurden so übernommen, wie sie schon in den Datenbanken vorliegen.

Tabelle 4 - Liste der untersuchten Referenzorganismen, die für die statistische Evaluierung des Programms herangezogen wurden.

Organismus	**Stamm**
Corynebacterium glutamicum	ATCC13032
Dinoroseobacter shibae	DFL-12
Escherichia coli	K-12
Pseudomonas aeruginosa	PAO1
Pseudomonas putida	KT2440
Sulfolobus solfataricus	P2
Thermus thermophilus	HB27
Yersinia pseudotuberculosis	IP-32953
Yersinia pseudotuberculosis	YPIII

2.10 Wahrheitsmaß

Für die statistische Auswertung der EnzymeDetector-Ergebnisse wurde ein Wahrheitsmaß benötigt, gegen das evaluiert werden konnte. Die Wahl fiel hier auf Swiss-Prot. Diese Datenbank enthält, wie bereits erwähnt, nur manuell überprüfte Daten von hoher Qualität. Ein weiteres Entscheidungskriterium war, dass in Swiss-Prot eine Vielzahl von verschiedenen Organismen abgedeckt ist. Oft kommen in anderen

Datenbanken dieser Qualität nur wenige oder sogar nur einzelne Organismen vor. Zur Definition eines Organismen-übergreifenden Wahrheitsmaßes sind diese somit nicht geeignet.

Einbezogen in die Evaluierung wurden jeweils nur Gene, die in Swiss-Prot annotiert waren. Dies umfasst im Durchschnitt nur 23 % der Gene, für die eine Annotation in einer der Annotationsquellen gefunden werden konnte. Über die anderen Gene kann leider keine definitive Aussage gemacht werden. Die Aussagen, die über die in Swiss-Prot annotierten Gene getroffen werden können, werden auch auf die anderen Gene der Organismen angewendet.

2.11 Ablauf EnzymeDetector

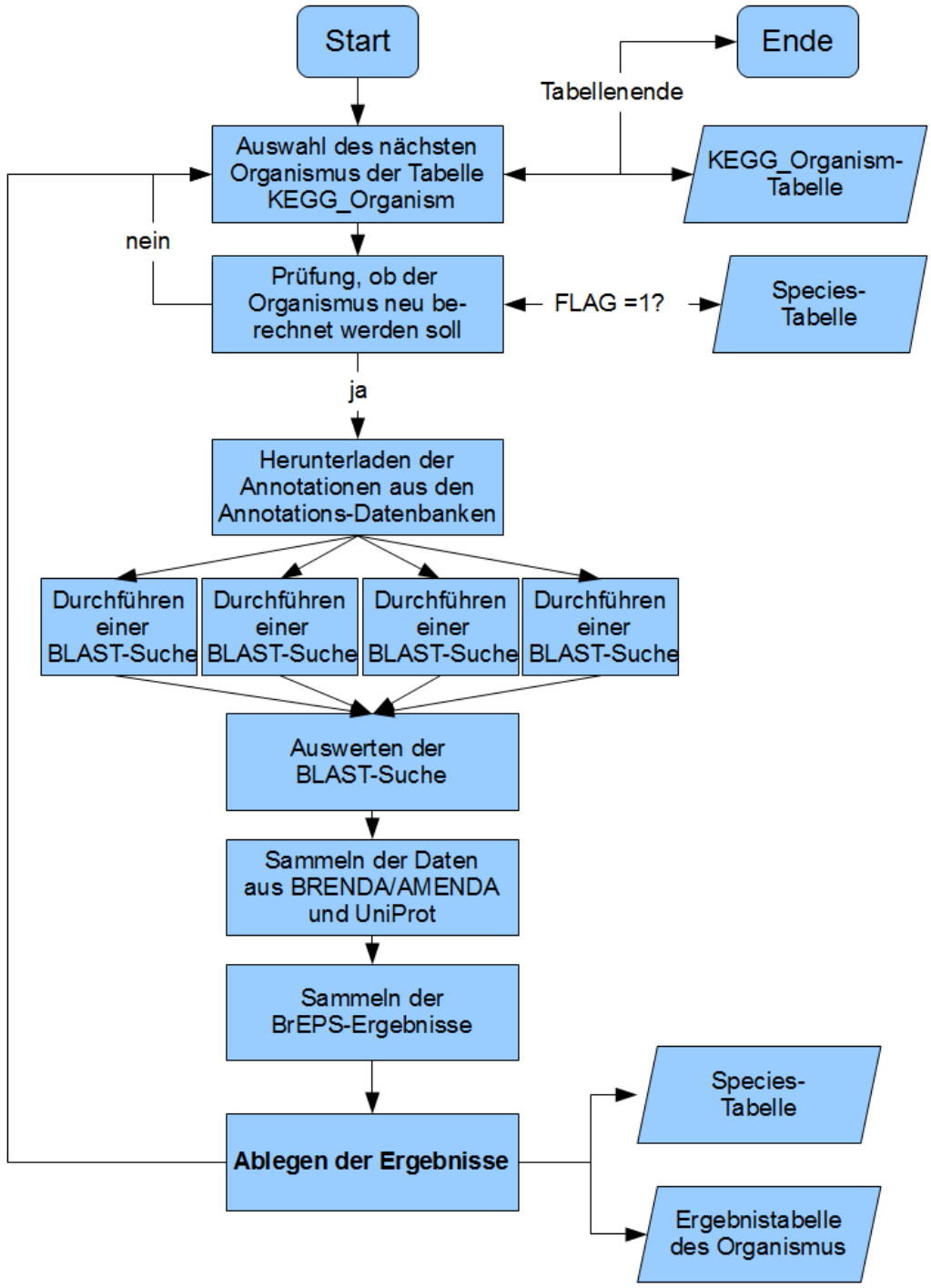


Abbildung 2 - Schematische Darstellung der Programmablaufs des EnzymeDetector mit den einzelnen Schritten, die vom Programm durchlaufen werden.

Der Ablauf des Programms EnzymeDetector ist in Abbildung 2 zu sehen. Dies ist der Ablauf des reinen Programms, nachdem schon die erforderlichen Datenbanken und Tabellen aus Punkt 2.6 angelegt bzw. aktualisiert wurden. Teile des Programms werden parallelisiert auf einem Rechencluster ausgeführt, damit ein Update jedes halbe Jahr realisierbar ist. Um die Gesamtrechenzeit weiter zu verkürzen, werden zusätzlich mehrere Instanzen des Programms auf verschiedenen Knoten des Clusters gestartet. Genauere Angaben zu den Spezifikationen des Clusters sind unter2.14zu finden.

2.11.1 Programmiersprache

Das gesamte Programm wurde in der Sprache Python entwickelt. Hierbei handelt es sich um eine Interpretersprache. Das bedeutet, dass der Quellcode nicht in eine ausführbare Datei umgewandelt wird, sondern zur Laufzeit eingelesen und interpretiert wird. Dies sorgt zwar dafür, dass Python-Programme systemunabhängig laufen, allerdings muss dafür eine längere Laufzeit in Kauf genommen werden. Neben der Standardbibliothek wurden weitere Module zur Erweiterung der Möglichkeiten eingebunden. Das wichtigste Modul, das für die Arbeit eine Rolle spielte, ist das Modul mysqlDB.

MySQLdb bietet eine Python-Datenbank-API, was eine leichte Kommunikation mit einem SQL-Datenbankserver ermöglicht[51]. Durch das Modul und die hergestellte Verbindung zum gewünschten SQL-Server können SQL-Statements einfach übergeben werden und die Rückantwort des Servers abgefangen werden. Dies war von großer Bedeutung für die vorliegende Arbeit, da alle Zwischenergebnisse, Endergebnisse und Berechnungsdaten des EnzymeDetectors in SQL-Tabellen abgelegt wurden.

2.11.2 Auswahl des Organismus

Nach dem Start der automatischen Berechnung des EnzymeDetectors muss zunächst ein Organismus ausgewählt werden, der prozessiert werden soll. Die Auswahl dieses Organismus geschieht mit Hilfe der in Punkt 2.5.1 beschriebenen Tabellen. Es wird auf die Tabelle „KEGG_Organism“ zugegriffen, um eine aktuelle Liste aller Organismen zu erhalten, die für eine Berechnung in Frage kommen. Anschließend wird in der Tabelle „Species“ für jeden dieser Organismen geprüft, ob dieser für eine Neuberechnung

markiert ist. Dies ermöglicht es einerseits, nur bestimmte Organismen für die Neuberechnung auszuwählen, und andererseits, eine parallele Ausführung des Programms auf mehreren Clusterknoten zu realisieren. Sobald mit der Berechnung eines Organismus begonnen wird, wird dies in der Tabelle vermerkt. Somit ist gesichert, dass kein anderer Prozess den gleichen Organismus berechnet. Das Datenbankmanagementsystem gewährleistet hierbei den sicheren Zugriff durch mehrere Prozesse und damit einen konsistenten Zustand der Datenbank.

Ist der aktuell ausgewählte Organismus nicht für eine Neuberechnung markiert bzw. wird/wurde dieser schon von einem anderen Programmprozess berechnet, wird die Liste weiter durchlaufen und der nächste Organismus geprüft.

2.11.3 Ermittlung des Genoms und der Datenbank-Annotationen

Wurde ein Organismus ermittelt, für den eine Berechnung durchgeführt werden soll, wird in einem ersten Schritt das Genom des Organismus mit den zugehörigen Annotationen automatisiert von den Annotationsdatenbanken (siehe2.8) abgefragt. Aus den Dateien werden die Geninformationen und ggf. ihre Enzymannotationen extrahiert und zwischengespeichert.

2.11.4 Durchführung einer BLAST-Suche

Anschließend wird für jedes Gen des Organismus eine BLAST-Suche (siehe2.3) durchgeführt. Um eine schnellere Berechnung zu ermöglichen und die Rechenkapazitäten des Clusters auszuschöpfen, werden immer vier Genabschnitte gleichzeitig mit der Datenbank verglichen. Die Ergebnisse der BLAST-Suchen aller Genabschnitte werden in einer SQL-Tabelle zwischengespeichert. So ist es einerseits möglich, auch unabhängig vom Programm die BLAST-Ergebnisse zu untersuchen, und andererseits wirkt sich dies sehr positiv auf die Laufzeit des Programms aus, da sonst sehr große Datenmengen zwischen der Funktionen übergeben werden müssten.

2.11.5 Auswertung der BLAST-Suche

Liegen die BLAST-Ergebnisse für das komplette Genom des Organismus vor, werden diese ausgewertet. Dieser Schritt entscheidet darüber, welche Enzyme in die BLAST-basierte („eigene“) Genomannotation übernommen werden.

Bei der Auswertung der BLAST-Suche kommen drei Kriterien zum Einsatz:

1. Die Vollständigkeit der EC-Nummern
2. Der E-value der BLAST-Ergebnisse
3. Die Häufigkeit des Ergebnisses in den BLAST-Ergebnissen

Für den Fall, dass es vollständige EC-Nummern gibt, d.h. solche, die mit allen vier Stellen definiert sind, werden unvollständige EC-Nummern ignoriert. Diese Prüfung wird für jeden Locus einzeln durchgeführt. Dieses Vorgehen wurde gewählt, da das Hauptaugenmerk der Funktionsfindung auf der anschließenden Modellierung des Organismus liegt. Für die Erstellung eines Modells ist die Substratspezifität von großer Bedeutung. Daher werden die vollständigen EC-Nummern auf Grund ihres größeren Informationsgehalts bevorzugt. Für den Fall, dass es nur unvollständige EC-Nummern unter den Lösungen gibt, werden diese als mögliche Ergebnisse übernommen, und die weiteren Kriterien werden für die Ergebnisse überprüft.

Neben der Vollständigkeit der EC-Nummern ist der E-value der BLAST-Ergebnisse Hauptentscheidungskriterium. Für jede EC-Nummer wird jeweils nur der kleinste E-value, mit dem sie unter den BLAST-Ergebnissen vorkommt, für die Entscheidung herangezogen.

Durch Vergleich der kleinsten E-values der BLAST-Ergebnisse der neun Referenzorganismen mit den vorhandenen Swiss-Prot-Annotationen wurde ein Bereich festgelegt, in dem der E-value der BLAST-Ergebnisse sich befinden muss, damit die zugehörigen EC-Nummern als Annotationskandidaten berücksichtigt werden. Dieser Bereich wird im Weiteren Signifikanzbereich genannt. Er umfasst alle Ergebnisse, deren E-values höchstens 30 Potenzen kleiner sind als der minimale E-value, mit dem ein Enzym für diesen Genabschnitt gefunden wurde.

Der Vergleich hat gezeigt, dass sich 30 Potenzen eignen, um ein ausgewogenes Verhältnis zwischen Abdeckung der Swiss-Prot-Enzyme und Nicht-Hinzufügen von ungewollten Enzymaktivitäten zu erhalten. Durchschnittlich lagen mindestens 95 % der Enzyme, die in der BLAST-Suche gefunden wurden und in Swiss-Prot annotiert waren, im Signifikanzbereich. Ferner liegen nur 17 % der Enzyme, die von BLAST zusätzlich gefunden wurden und nicht in Swiss-Prot annotiert waren, in diesem Bereich. Die Zahlen zeigen, dass dank der Wahl des Signifikanzbereiches der Großteil der gewünschten Enzymfunktionen abgedeckt ist, ohne dass viele nicht relevante Enzymfunktionen als Annotationskandidaten in Erwägung gezogen werden.

Findet sich im Signifikanzbereich nur eine EC-Nummer, wird diese als eindeutige Annotation für den Genabschnitt übernommen. Befinden sich mehrere EC-Nummern in diesem Bereich, werden alle als mögliche Kandidaten gewertet und an das nächste Entscheidungskriterium weitergegeben.

Als letztes Entscheidungskriterium wird, falls bis dahin keine eindeutige Entscheidung möglich war, die Häufigkeit des Auftretens der EC-Nummern unter allen BLAST-Hits ausgewertet. Als Ergebnis der BLAST-Suche erhält man eine Liste von Hits. Unter diesen Hits kann es natürlich mehrere homologe Sequenzen geben, hinter denen die gleiche EC-Nummer steht. Ausgewertet werden nur EC-Nummern, die im Signifikanzbereich des besten E-values liegen. Eine EC-Nummer wird dann als signifikant in der Häufigkeit ihres Auftretens betrachtet, wenn sie häufiger als fünfmal innerhalb des Signifikanzbereiches unter den BLAST-Ergebnissen vorkommt. Dieser Wert wurde nach manueller Auswertung der Datensätze festgelegt. Es hat sich gezeigt, dass eine höhere Grenze zum Ausschluss von für den Organismus entscheidenden Enzymfunktionen führt. Diese Grenze gilt nur, wenn es Ergebnisse gibt, die diese Grenze überschreiten. Gibt es ausschließlich EC-Nummern mit einer Häufigkeit kleiner als fünf, werden diese als signifikant gewertet. Diese Einschränkung sorgt dafür, dass BLAST-Ergebnisse von Proteinsequenzen, die (bis jetzt) kaum Homologe in der Datenbank haben, nicht aussortiert werden, sondern in diesem Fall als mögliche Kandidaten übernommen werden. Gibt es EC-Nummern mit einer Häufigkeit größer als fünf, werden diese denen mit einer geringeren Häufigkeit vorgezogen. So soll verhindert werden, dass EC-Nummern als Annotationskandidaten übernommen werden,

die zum Beispiel nur ein einziges Mal in den gesamten BLAST-Ergebnissen auftreten. Hier ist eher davon auszugehen, dass es sich um eine falsch annotierte Sequenz handelt.

Ist es möglich, mit Hilfe der drei Kriterien eine eindeutige Annotation für einen Genabschnitt zu ermitteln, wird diese als eindeutiges Ergebnis übernommen. Ist es nicht möglich, werden alle Annotationen als Kandidaten übernommen, denen eine vollständige EC-Nummer zugeordnet ist, die im Signifikanzbereich liegen und eine signifikante Häufigkeit aufweisen. Für das erste und das letzte Kriterium gelten die beschriebenen Ausnahmen.

2.11.6 Daten aus BRENDA/AMENDA und Swiss-Prot

Zusätzlich zu den Genomannotationen aus den Annotationsdatenbanken (NCBI/KEGG/PEDANT) wird des Weiteren ermittelt, welche Enzyme laut BRENDA und AMENDA im Organismus vorliegen. In BRENDA finden sich sowohl sequenzspezifische als auch -unspezifische Informationen. AMENDA beinhaltet nur sequenzunspezifische. Die sequenzspezifischen Annotationen werden zu den entsprechenden Genen als Information hinzugefügt. Die sequenzunspezifischen Annotationen werden genunabhängig gespeichert. Somit gehen gesicherte Informationen zum Stoffwechsel des Organismus nicht verloren, auch wenn sie nicht genspezifisch vorliegen. Diese Informationen sind von sehr großem Wert für Modellierer, für die zwar eine Enzymfunktion, nicht aber die Genzuordnung entscheidend ist.

Die Annotationsdaten aus Swiss-Prot werden aus der vorher angelegten Tabelle (siehe2.8.4) ausgelesen und ebenfalls den entsprechenden Genen zugeordnet.

2.11.7 BrEPS-Ergebnisse

Anschließend werden die BrEPS-Ergebnisse für den untersuchten Organismus ermittelt und den jeweiligen Genen zugeordnet. Hierfür werden die schon bestehenden BrEPS-Muster (zu Berechnung siehe Punkt 2.4 auf Seite 15) genutzt und mit den Genen des Organismus eine Suche durchgeführt. Durch die Vorberechnung der Muster dauert dieser Vorgang nur wenige Minuten pro Organismus.

2.11.8 Ablegen der Ergebnisse

Im Anschluss an das Sammeln und Erstellen der Annotationen wird eine SQL-Tabelle für die Speicherung der Ergebnisse angelegt. Zusätzlich werden die Informationen zum Organismus in der 'Species'-Tabelle aktualisiert. Es werden der Name der Ergebnis-Tabelle, die maximale Relevanz, die das Programm EnzymeDetector ermittelt hat und die Annotationsquellen, die für den Organismus eingebunden werden konnten, gespeichert.

In die Ergebnis-Tabelle wird pro ermittelter möglicher Annotation, d.h. Gen-Enzym-Kombination bzw. Enzym-Annotation ohne Geninformation, ein Eintrag angelegt. Die zugehörigen Informationen lassen sich in drei Klassen einteilen:

- die gen- und organismenspezifische Information: der Organismenname , Genstart und Genstop, der Locus Tag und die GI des Gens. (bei den genunspezifischen fällt ein Teil der Information weg)

- die BLAST-spezifische Information: die EC-Nummer des Enzyms, der empfohlene Name des Enzyms (Information aus BRENDA), die Position des BLAST-Hits, die Häufigkeit der EC-Nummer unter allen BLAST-Hits, der beste E-value, mit dem das Enzym in der gesamten UniProt-Datenbank gefunden wurde, der beste E-value, mit dem das Enzym in Swiss-Prot gefunden wurde und die Anzahl der Kandidaten, die für die BLAST-basierte Genomannotation übernommen wurden.

- die auswertungsspezifische Information: in diesem Block findet sich die Information, in welcher der Annotationsquellen das Enzym diesem Genabschnitt zugeordnet wurde, und die Gesamtrelevanz (siehe2.12auf Seite 34), die die Annotation dadurch erhält.

2.12 Relevanzwert des EnzymeDetectors

Jede Annotation, die für einen Organismus gefunden wurde, bekommt einen Relevanzwert zugeordnet. Dieser setzt sich aus der Summe der Einzelrelevanzen der Annotationsquellen zusammen. Der höchste erreichbare Relevanzwert für einen

Organismus ist abhängig davon, welche Annotationsquellen eingebunden werden konnten.

Die Einzelrelevanz, die einer Annotation zugeordnet wurde, wurde mit Hilfe des durchschnittlichen F1-Maßes der Quelle im Vergleich mit Swiss-Prot bestimmt. Hierfür wurden die Swiss-Prot-Annotationen der Gene, bei denen eine Annotation in Swiss-Prot vorhanden war, mit den Annotationen aus der jeweiligen Annotationsquelle verglichen und das F1-Maß bestimmt. Ermittelt wurde jeweils das durchschnittliche F1-Maß der neun Referenzorganismen.

Das F1-Maß setzt sich aus der Trefferquote (dem sogenannten *Recall*) und der Genauigkeit (genannt *Precision* zusammen, wobei beide Werte gleich gewichtet werden.

Die Formeln zur Berechnung der einzelnen Werte sind im Folgenden aufgeführt:

$$recall = 100 \cdot (TP / (TP + FN))$$

$$precision = 100 \cdot (TP / (TP + FP))$$

$$F1 = 2 \cdot (precision \cdot recall) / (precision + recall)$$

TP steht dabei für die richtig Positiven, also die Annotationen, die in Swiss-Prot vorlagen und auch in der entsprechenden Quelle gefunden wurden. FN steht für die falsch Negativen, also die Annotationen, die zwar in Swiss-Prot annotiert waren, nicht aber in der untersuchten Quelle. Daneben steht FP für die falsch Positiven, d.h. die Annotationen, die zwar in der jeweiligen Datenbank auftraten, aber nicht durch Swiss-Prot bestätigt wurden.

Die Summe der Einzelrelevanzen, die für die verschiedenen Annotationsquellen ermittelt wird und die für eine Annotation vergeben wird, falls die entsprechende Datenbank diese beinhaltet, ergibt die Gesamtrelevanz einer Annotation. Diese steht für die Güte selbiger und hilft die Annotationen zu bewerten und miteinander zu vergleichen.

2.13 Operonanalyse

Zusätzlich zu den Informationen, die aus den Annotationsquellen und der BLAST-Auswertung in die Ergebnisse des EnzymeDetectors eingehen, können optional die Informationen einer Operonanalyse in die Ergebniswertung einbezogen werden.

2.13.1 Ablauf der Operonanalyse

Diese Analyse wird für die gespeicherten Ergebnisse des Programms EnzymeDetector auf Basis der Daten aus der Operon-Datenbank DOOR durchgeführt. Wie schon unter Punkt 2.2 (Seite 13) beschrieben, geht man davon aus, dass die Gene eines Operons einen funktionellen Zusammenhang haben. Somit ist die Wahrscheinlichkeit hoch, dass die Enzyme eines Operons in einem gemeinsamen Stoffwechselweg auftreten. Diese Tatsache macht man sich bei der Operonanalyse zunutze, um Hinweise zu erhalten, welche der gefundenen Enzymkandidaten am wahrscheinlichsten sind.

Ein Beispiel für den Ablauf der Operonanalyse ist in Abbildung 3 zu sehen. Als erster Schritt der Operonanalyse wird ermittelt, welche Gene des Organismus zu einem Operon gehören. Im vorliegenden Beispiel besteht das Operon mit der Operon-ID 55 aus *E. coli* aus zwei Genen. Für diese Gene werden anschließend die Enzym-Kandidaten aus der EnzymeDetector-Ergebnistabelle ermittelt. Für das eine Gen wurde ein eindeutiger Enzym-Kandidat gefunden, und für das andere werden drei verschiedene Enzyme vorgeschlagen.

Diese Enzym-Kandidaten werden daraufhin überprüft, ob sie in einem gemeinsamen Stoffwechselweg auftreten. Es ist nur entscheidend, ob Kandidaten von verschiedenen Loci in einem Stoffwechselweg vorkommen. Allerdings ist es möglich, dass die Kandidaten zwar auf dem gleichen Gen liegen, sich aber trotzdem nicht überschneiden. In diesem Fall werden sie gleichwertig betrachtet.

Gewertet werden nur Ergebnisse, bei denen die Schnittmenge der Enzyme des Stoffwechselwegs mit den Enzym-Kandidaten des Operons größer als eins ist. Die Information, aus welchen Enzymen ein Stoffwechselweg besteht, wird aus den beiden Stoffwechselweg-Datenbanken KEGG und MetaCyc entnommen. Beim gezeigten Beispiel kommen zwei der drei vorgeschlagenen Enzym-Kandidaten in der Acetyl-

CoA-Biosynthese vor. Im vorliegenden Fall erhalten diese beiden Enzyme einen Positiv-Score, der als Vermerk in der Ergebnistabelle des EnzymeDetectors abgelegt wird.

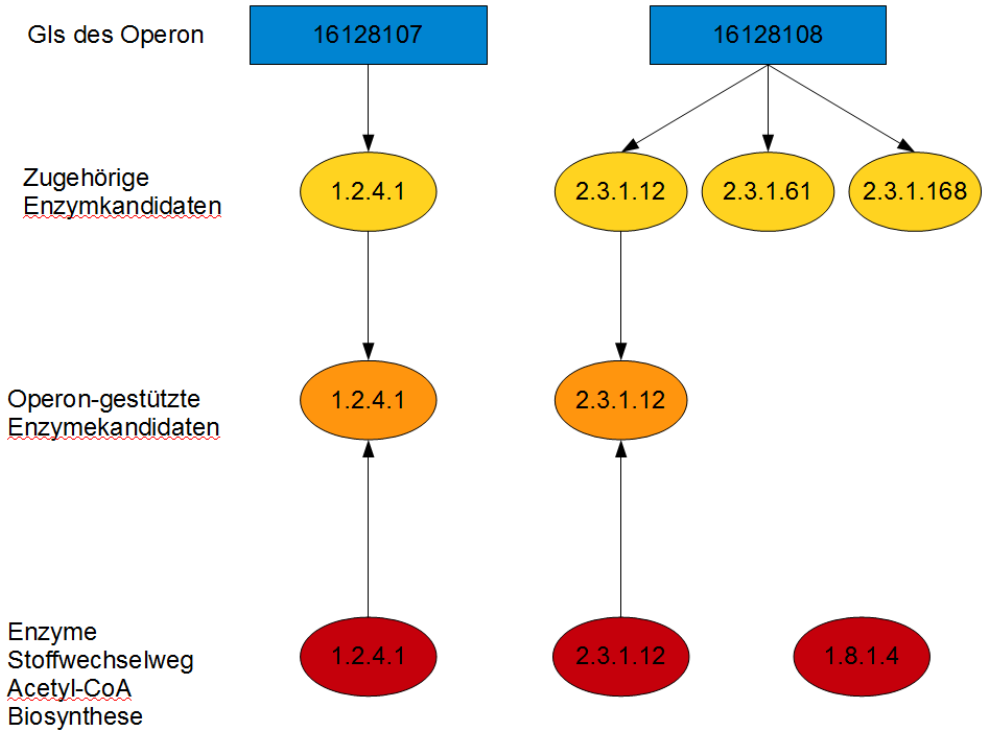


Abbildung 3 - Operonanalyse anhand eines Beispiels aus E. coli

2.13.2 Score der Operonanalyse

Wie schon beschrieben, erhalten die Enzyme eines Operons, die in einem gemeinsamen Stoffwechselweg vorkommen, einen Positiv-Score. Der Positiv-Score ist abhängig davon, wie viele Enzyme aus dem Operon in dem Stoffwechselweg vorkommen. Kommen nur zwei Enzyme aus dem Operon in dem Stoffwechselweg vor, erhalten beide jeweils den Score 2, sind drei Enzyme des Operons vertreten, erhalten diese den Score 3 usw.

Für den Fall, dass für ein Gen mehrere Enzymkandidaten gefunden werden, die in einem Stoffwechselweg vorkommen, wird ermittelt, ob sich die Positionen der BLAST-Treffer dieser Enzyme überschneiden, es also wirkliche Konkurrenten sind, oder ob sie nebeneinander liegen, also beide unabhängig auf dem Gen vorhanden sein können. Im

letzteren Fall erhalten beide Enzyme den vollen Score. Im ersten Fall wird der Score durch die Anzahl der konkurrierenden Enzyme geteilt.

In Abbildung 4 ist ein Beispiel für das Vorkommen mehrerer Enzym-Kandidaten für ein Gen gezeigt. Enzym 1 hat in diesem Fall keinen Konkurrenten und erhält somit den vollen Score. Die Enzyme 2, 3 und 4 überschneiden sich allerdings in ihrer Lage und erhalten somit jeweils nur ein Drittel des eigentlichen Scores.

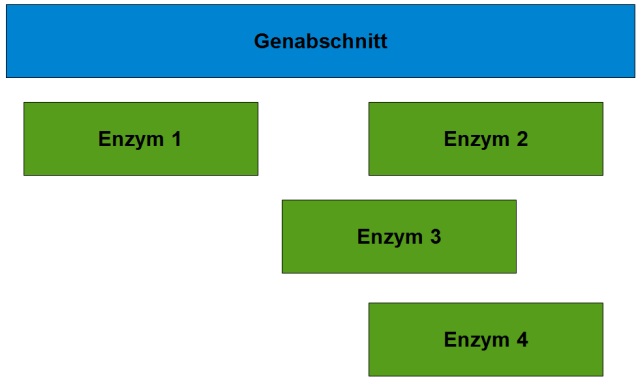


Abbildung 4 - Beispiel für die Einbeziehung der Position der BLAST-Treffer in die Auswertung der Operonanalyse

Der Score wird für jeden Stoffwechselweg, in dem die Enzym-Kandidaten vorkommen, aufaddiert. Abschließend wird er dann gemittelt, d.h. durch die Zahl der Stoffwechselwege, in denen das Enzym vorkommt, geteilt.

Kommt also ein Enzym eines Operons in einem Stoffwechselweg gemeinsam mit zwei weiteren Enzymen des Operons vor, erhält es für diesen Stoffwechselweg den Score 3.

In einem weiteren Weg kommt es gemeinsam mit nur einem anderen Enzym vor und erhielte somit den Score 2. Allerdings hat es für diesen Weg einen Konkurrenten auf dem gleichen Locus. Somit wird der eigentlich Score für diesen Stoffwechselweg durch die zwei Konkurrenten geteilt und beide erhalten den Score 1.

Die Summe der Scores der einzelnen Stoffwechselwege wird abschließend durch die Anzahl der Wege, in diesem Beispiel zwei, geteilt und somit erhält das Beispielenzym einen Gesamtscore von 2.

Der gemittelte Score wird in die EnzymeDetector Ergebnistabelle als Vermerk eingetragen.

2.13.3 Wertung der Annotationen

Da, wie im Ergebnisteil dieser Arbeit gezeigt wird, die Daten nicht signifikant genug für eine eigenständige Genomannotation sind, werden sie nicht in die Berechnung des Relevanzwertes des EnzymeDetector einbezogen. Allerdings werden sie ausgewertet und als zusätzliche Information in die EnzymeDetector-Ergebnistabellen eingefügt, um in unklaren Fällen als Entscheidungshilfe zu dienen.

2.14 Rechencluster

Die Berechnungen des Programms EnzymeDetector wurden, wie bereits erwähnt, auf einen Rechencluster ausgelagert. Dies war nötig, um eine angemessene Rechenzeit und Performance zu erhalten.

Tabelle 5 - Spezifikationen der einzelnen Cluster-Knoten mit ihrer Rechenleistung

Knoten	**Arbeitsspeicher**	**CPU**		**Anzahl Kerne**
Masterknoten	6 x 4 GB	2 x Intel® Xeon® CPU	2.40 GHz	
neuer Rechenknoten	6 x 2GB	2 x Intel® Xeon® CPU	2.40 GHz	je 4 echte und 4 virtuelle Kerne → 16 Kerne insgesamt
alter Rechenknoten	8 x 2 GB	2 x Dual-Core AMD Opteron™		je 2 echte Kerne → 4 Kerne

Der Cluster besteht aus einem Master- und zehn leistungsstärkeren Rechenknoten sowie acht nicht so leistungsstarken Rechenknoten. In Tabelle 5 sind die Spezifikationen der genutzten Knoten aufgelistet. Für ein Update des Programms EnzymeDetector werden die zehn leistungsstärkeren Rechenknoten komplett ausgelastet. Es wird pro Knoten ein

Organismus berechnet, wobei die BLAST-Suche parallelisiert in 4 Prozessen, die jeweils 4 Prozessorkerne belegen, durchgeführt wird.

2.15 EnzymeDetector Webinterface

Für die Ergebnisse aus den EnzymeDetector-Berechnungen wurde ein Webinterface geschaffen, welches einen einfachen öffentlichen Zugang zu den Daten für alle ermöglicht. Dieses ist unter http://enzymedetector.tu-bs.de frei verfügbar. Die Daten der Operonanalyse stehen auf dem Webinterface noch nicht zur Verfügung. Sie werden aber in naher Zukunft eingebaut werden.

2.15.1 genutzte Programmiersprachen

Für die Realisierung des Webinterface wurde serverseitig mit der Programmiersprache Python gearbeitet. Dies bot sich an, da auch das EnzymeDetector Programm in dieser Sprache entwickelt wurde. Clientseitig wurde mit HTML und für die Dynamik mit JavaScript gearbeitet. Das Aussehen der Seite wurde mittels Cascading Style Sheets (CSS) festgelegt.

Bei der Umsetzung der Seite wurde das Hauptaugenmerk neben dem einfachen Zugang zu den Daten auf die Dynamik gelegt. Die Seiten selbst wurden nur als Hüllen programmiert. Diese werden mittels SQL-Abfragen mit den Daten aus den dafür angelegten Datenbanktabellen (Punkt 2.5.2) gefüllt.

Auf diese Weise kann der Inhaltsumfang der Webseite ohne Änderung des Webseiten-Codes leicht geändert werden.

2.16 Molfiles

Molfiles sind Dateien, die Informationen über chemische Verbindungen enthalten. Es handelt sich um ein von Molecular Design Limited (MDL) entwickeltes Format, das für die textbasierte tabellarische Speicherung von Informationen zu einer chemischen Struktur entwickelt wurde, um den Datenaustausch zwischen verschiedenen Software-Programmen zu erleichtern. Die gespeicherten Informationen umfassen die Atome des

Moleküls mit den entsprechenden Bindungen, der Konnektivität und den Koordinaten der Atome. Ein Molfile besteht aus zwei klar definierten Blöcken: Zum einen aus dem dreizeiligen Headerblock, der allgemeine Informationen beinhaltet wie den Dateinamen, Informationen zum Editor, mit dem die Struktur erstellt wurde, etc.. Zum anderen besteht es aus der Connection Table, die wiederum aus mehreren Blöcken besteht, der Zählerlinie, dem Atomblock, dem Bindungsblock und optional dem Eigenschaften-Block.

Die Zählerlinie enthält Informationen zu Atomanzahl, Bindungsanzahl etc.. Im Atomblock, der jedes Atom durch eine Zeile repräsentiert, sind für jedes Atom die Koordinaten, das Atomsymbol und weiterführende Angabe zum Atom zu finden. Der Bindungsblock enthält Informationen zu den Bindungen, die im Molekül zwischen den einzelnen Atomen vorkommen. Der optionale Eigenschaftsblock ist in vielen Molfiles nicht vorhanden. Dieser speichert zusätzliche Informationen zur Struktur, zur Stereochemie, zur 3D-Geometrie oder ähnlichem. Diese Informationen ergeben sich aus den 3D-Koordinaten der Atome und der Bindungstopologie. Jedes Molfile endet mit dem Kürzel 'M END'.[52]

Für den in der Arbeit erweiterten Atom Mapping- Algorithmus werden diese Molfiles eingelesen und die Informationen zu den Atomen und Bindungen des Moleküls genutzt. So können auch nur Reaktionen mit dem Algorithmus untersucht werden, bei denen für alle Reaktionspartner die entsprechenden Molfiles vorliegen.

2.17 „Atom Mapping“-Algorithmus

Um den Fluss der Atome ganzer Reaktionen nachvollziehen zu können, wurde ein bestehendes Programm von Markus Leber[26] erweitert. Bei dem Programm handelt es sich um einen „Maximal Common Subgraph“(MCS)-Algorithmus. Ziel ist es, die maximale gemeinsame Teilstruktur zweier Moleküle mit Hilfe von Graphalgorithmen zu ermitteln. In dem entwickelten Algorithmus werden zwei bekannte Graphalgorithmen miteinander kombiniert, um ein möglichst gutes Ergebnis zu erhalten.

2.17.1 Bron-Kerbosch-Algorithmus

Für den Bron-Kerbosch-Algorithmus werden die beiden zu untersuchenden Moleküle in Graphen überführt. Hierbei bilden die Atome der Moleküle die Graphknoten und die Bindungen zwischen den Atomen die Graphkanten. Anschließend wird aus den beiden Molekülgraphen ein gemeinsamer Produktgraph erstellt. Für jedes Atom, das in beiden Molekülgraphen gleich ist, d.h. das gleiche Atomsymbol hat, wird ein Knoten des Produktgraphen angelegt. Die Kanten des Produktgraphen hängen davon ab, ob die Atome, die in den Produktknoten enthalten sind, in den Ursprungsgraphen verbunden sind oder nicht.

Nach der Erstellung des Produktgraphen wird in einem weiteren Schritt des Algorithmus nach maximalen Cliquen gesucht. Eine Clique innerhalb eines Graphen ist eine Teilmenge seiner Knoten, innerhalb derer alle Knoten paarweise miteinander verbunden sind. Bei dem Algorithmus handelt es sich um ein Backtracking-Verfahren. Das bedeutet, dass der Algorithmus versucht, die Clique schrittweise möglichst weit auszudehnen. Ist klar, dass die aktuelle Teillösung nicht zu einer Gesamtlösung führen kann, springt der Algorithmus einen oder mehrere Schritte zurück und versucht von dort aus, auf einem alternativen Weg zu einer Gesamtlösung zu kommen. So werden alle Möglichkeiten nach und nach getestet und die optimale Lösung gefunden.

2.17.2 McGregor-Algorithmus

Konnten durch den Bron-Kerbosch-Algorithmus nicht alle Atome der beiden Moleküle aufeinander abgebildet werden, wird der McGregor-Algorithmus für die übrig-gebliebenen Atome angewendet.

Es werden alle schon verknüpften Atome durchlaufen und ihre Bindungspartner ermittelt. Für diese Bindungspartner wird jeweils geprüft, ob auch diese schon verknüpft sind. Ist dies nicht der Fall, wird geprüft, ob es an der gleichen Stelle im Partnermolekül ebenfalls einen nicht verknüpften Bindungspartner mit dem gleichen Atomsymbol gibt. Ist dies der Fall, werden die beiden Atome zur Clique hinzugefügt und die Clique so erweitert.

Auf diesem Weg erhält man durch den kombinierten Algorithmus die Information, welches Atom des einen Moleküls zu welchem Atom des anderen Moleküls zugeordnet werden kann. Wichtig ist zu erwähnen, dass bei dem Algorithmus nur die Atomsymbole eine Rolle spielen. Die Art der Bindung zwischen den Atomen, ob zum Beispiel eine Einfach- oder eine Doppelbindung vorliegt, wird nicht beachtet.

Die Moleküle werden in Form von Molfiles eingelesen und verglichen.

2.18 BRIME

Der BRaunschweig Interactive Metabolism Explorer (BRIME)[53] ist ein Tool, mit dem metabolische Netzwerke dargestellt, bearbeitet und durchsucht werden können. Die Abbildungen der generischen Stoffwechselkarte, die in dieser Arbeit verwendet werden, sind teilweise mit diesem Tool erstellt worden.

2.19 Cytoscape

Bei Cytoscape[54] handelt es sich um ein Open-Source-Programm zum Visualisieren komplexer Netzwerke. Es ist allerdings nicht speziell für die Visualisierung metabolischer Netzwerke gemacht. Das Hauptaugenmerk liegt bei diesem Programm darin, dass beliebige Arten von Daten als Attribute zu einem Knoten bzw. einer Kante hinzugefügt werden können.

Aus diesem Grund wurde Cytoscape im Rahmen dieser Arbeit auch genutzt, um die generische Stoffwechselkarte (siehe2.22) zu erstellen. Es war das einzige der in Frage kommenden Programme, das es ermöglicht, Knoten und Kanten eigene Attribute zuzuweisen. Dies war entscheidend für die vorliegende Arbeit.

Des Weiteren ist es möglich, den generellen Funktionsumfang von Cytoscape durch eine Vielzahl von frei verfügbaren Plugins zu erweitern bzw. selbst Plugins zu schreiben.

Cytoscape bietet eine große Auswahl an verschiedenen Eingabe- und Ausgabeformaten, wie zum Beispiel SBML, GML, XGMML, PSI-MI. Die generische Stoffwechselkarte

wurde im Ausgabeformat XGMML erstellt, da dieses Dateiformat leicht zu bearbeiten ist.

Cytoscape wurde des Weiteren genutzt, um einige der Abbildungen der Stoffwechselkarte, die in dieser Arbeit verwendet wurden, zu erstellen.

2.20 XGMML

Die eXtensible Graph Markup and Modeling Language (XGMML)[55] ist eine Erweiterung von XML basierend auf der Graph Modeling Language (GML). Sie ermöglicht einen einfachen Austausch von Graphen zwischen verschiedenen Programmen und Autoren. Neben den Netzwerkinformationen enthält das XGMML-Format Knoten-, Kanten- und Netzwerkattribute. Gleichzeitig bietet XGMML die Flexibilität, die mit jedem XML-Format einhergeht.

Die erstellte Stoffwechselkarte wurde mit Hilfe von Cytoscape im XGMML-Format angelegt. Alle Informationen der Karte sind in der XGMML-Datei gespeichert, die dann sowohl mit Hilfe verschiedener Grapheditoren angezeigt werden können, als auch einfach manipuliert werden können.

2.21 DOM-Parser

Mit Hilfe des Python-Moduls xml.dom.minidom[56] lassen sich auf sehr einfache Art und Weise XML-Formate auslesen und manipulieren. Da es sich bei der XGMML-Datei der Stoffwechselkarte um ein XML-Format handelt, wurde der DOM-Parser genutzt, um die Karte einerseits in eine SQL-Tabelle umzuschreiben und andererseits verschiedene Grafikfunktionen durch eine Manipulation der XGMML-Datei umzusetzen. Dank des Python-Moduls kann auf die Datei entsprechend ihrer XML-Strukturen zugegriffen werden. Ein zeilenweises Auslesen ist so überflüssig.

Für die grafische Manipulation der XGMML-Datei wurden Knoten- und Kanten-Attribute in der Datei manipuliert bzw. Knoten und Kanten teilweise komplett gelöscht und die resultierende XGMML-Datei abgespeichert. Diese kann mit Programmen wie Cytoscape eingelesen und dargestellt werden.

2.22 Stoffwechselkarte

Im Rahmen der vorliegenden Arbeit wurde an der Planung und Erstellung einer generischen Stoffwechselkarte mitgearbeitet. Ziel dieses Projektes war es, eine übersichtliche, intuitive und dynamische Karte eines Stoffwechselnetzwerks herzustellen.

Die Karte wurde mit Hilfe des Zeichenprogramms Cytoscape erstellt. Für die Erstellung der Karte wurden die Stoffwechselkarten von KEGG und MetaCyc zur Orientierung hinzugezogen und durch fehlende Reaktionen erweitert. Zudem erfolgte eine Korrektur der Unstimmigkeiten, die in diesen Karten teilweise vorhanden waren. Diese Arbeiten erfolgten teilweise im Rahmen von in der Arbeit betreuten Praktika oder durch Werksvertrag-Studenten.

Der Stoffwechselgraph wurde als bipartiter Graph repräsentiert. Das bedeutet, dass es zwei eindeutig bestimmbare Teilmengen unter den Knoten gibt, so dass es keine Kante gibt, die zwei Knoten innerhalb ein und derselben Teilmenge verbindet. Die zwei Teilmengen der Knoten stellen Enzyme und Metabolite dar. Der gesamte Graph wurde per Hand angeordnet, wobei darauf geachtet wurde, dass die Anordnung der einzelnen Stoffwechselwege den bekannten Anordnungen aus Lehrbüchern entsprach. Diese Maßnahme erleichtert das Zurechtfinden innerhalb der Karte. Außerdem wurde der Graph in vier Quadranten aufgeteilt, die um den Zentralstoffwechsel, bestehend aus Pentosephosphatweg, Glykolyse, Citratzyklus und Gluconeogenese, angeordnet wurden. Das Gesamtbild der erstellten Stoffwechselkarte ist in Abbildung 5 zu sehen. Diese Abbildung bietet nur einen groben Überblick. Um Details erkennen zu können, ist eine höhere Auflösung, ein größeres Format bzw. ein kleinerer Ausschnitt der Karte nötig.

Im 1.Quadranten (links oben) befinden sich die Stoffwechselwege der Mono-, Di- und Polysaccharide und deren Derivate, die verschiedenen Gärungsarten, die Photosynthese. der C4/CAM-Stoffwechsel und der Peptidoglykan-Stoffwechsel.

Im 2. Quadranten (rechts oben) befinden sich der Metabolismus der Coenzyme, Cofaktoren und Vitamine und der Häm-Stoffwechsel sowie die Stoffwechsel der Pyrimidin- und Purinderivate.

Im 3. Quadranten (links unten) finden sich Fettsäure-Stoffwechsel, der Stoffwechsel von Cholesterol und seinen Derivaten, der Lipid- und Sphingolipid-Stoffwechsel und die Biosynthese der Isoprenoide.

Im 4. Quadranten (rechts unten) findet sich der Metabolismus der Aminosäuren und deren Derivate. Außerdem umfasst dieser Quadrant die Stickstofffixierung, die Nitratreduktion, den Harnstoffzyklus und die Sulfatassimilation.

Abbildung 5 - generische Stoffwechselkarte, die im Rahmen der Arbeit mit geplant und betreut wurde [Darstellung BRIME]

Es wurde jeweils darauf geachtet, dass Stoffwechselwege, die unter den Organismen nicht häufig vorkommen, eher am Rand des Netzwerkes positioniert werden. Dies macht die Ansichtsoption, nur organismenspezifische Stoffwechselwege anzeigen zu lassen, übersichtlicher und so entstehen keine großen Lücken mitten im Netzwerk.

Wie erwähnt gibt es zwei Knotentypen: die Enzyme und die Metabolite. Die Metabolite wurden mit Hilfe eines weiteren Attributs in Haupt- und Nebenmetabolite unterschieden, die auch durch unterschiedliche Knotenformen repräsentiert werden. Hauptmetabolite sollen nach Möglichkeit nur einmal im Gesamtnetzwerk auftreten. Dies war leider nicht immer realisierbar. Nahmen Metabolite an Reaktionen teil, die in der Karte sehr weit auseinanderlagen, wurden sie mehrfach in der Karte angelegt. Nebenmetabolite wie NADH und ATP wurden für jede Reaktion extra angelegt. Wären auch diese Metabolite nur in wenigen Instanzen realisiert worden, hätte die Übersichtlichkeit der Karte sehr gelitten, da diese Nebenmetabolite an sehr vielen Reaktionen des Netzwerkes teilnehmen.

Neben den Metabolit- bzw. Enzymnamen wurden für alle Knoten zusätzliche Informationen in selbstkreierten Attributen abgelegt. Diese Informationen umfassen den Stoffwechselweg, dem der Knoten zugeordnet ist, sowie den Hauptstoffwechselweg des Knotens, falls er in mehreren Stoffwechselwegen vorkommt. Des Weiteren wurde optional gespeichert, in welcher Karte in KEGG bzw. MetaCyc die vorliegende Reaktion zu finden ist. Dies erleichtert es nachzuvollziehen, aus welcher Quelle die dargestellte Information stammt, und Fehlerquellen einzugrenzen.

Für Knoten des Typs Enzym wurde außerdem die EC-Nummer und die Information abgelegt, ob es sich um eine reversible Reaktion handelt. Für Metabolite wurde zusätzlich zum Namen, da dieser kein eindeutiger Bezeichner ist, noch die Ligand-ID aus BRENDA abgelegt.

Die gesamte Karte wurde, wie unter Punkt 2.5.3 erwähnt, auch in einer relationalen SQL-Datenbank gespeichert, die alle Informationen enthält, die per Hand angelegt wurden. Auch erleichtert das Umschreiben der Informationen in eine Datenbank die Manipulation der Daten im Nachhinein, falls sich zum Beispiel eine Ligand-ID ändert o.Ä.. Diese Arbeit erfolgte in einem in der Arbeit mitbetreuten Etagenpraktikum und wurde nachfolgend durch Melanie Busch weiterbearbeitet und fertiggestellt.

3 Ergebnisse und Diskussion

3.1 EnzymeDetector

Für die statistische Auswertung des Programms EnzymeDetector wurden die neun Referenzorganismen, die in Tabelle 4 (Seite 26) aufgelistet sind, explizit betrachtet. Des Weiteren wurden Durchschnittswerte über alle 90 untersuchten Organismen angegeben (Liste dieser Organismen zu finden unter Anhang 6). Die Daten aus der Operonanalyse fließen in diese Statistik nicht mit ein. Sie werden unter Punkt 3.2 gesondert betrachtet.

3.1.1 Relevanzwerte der einzelnen Annotationsquellen

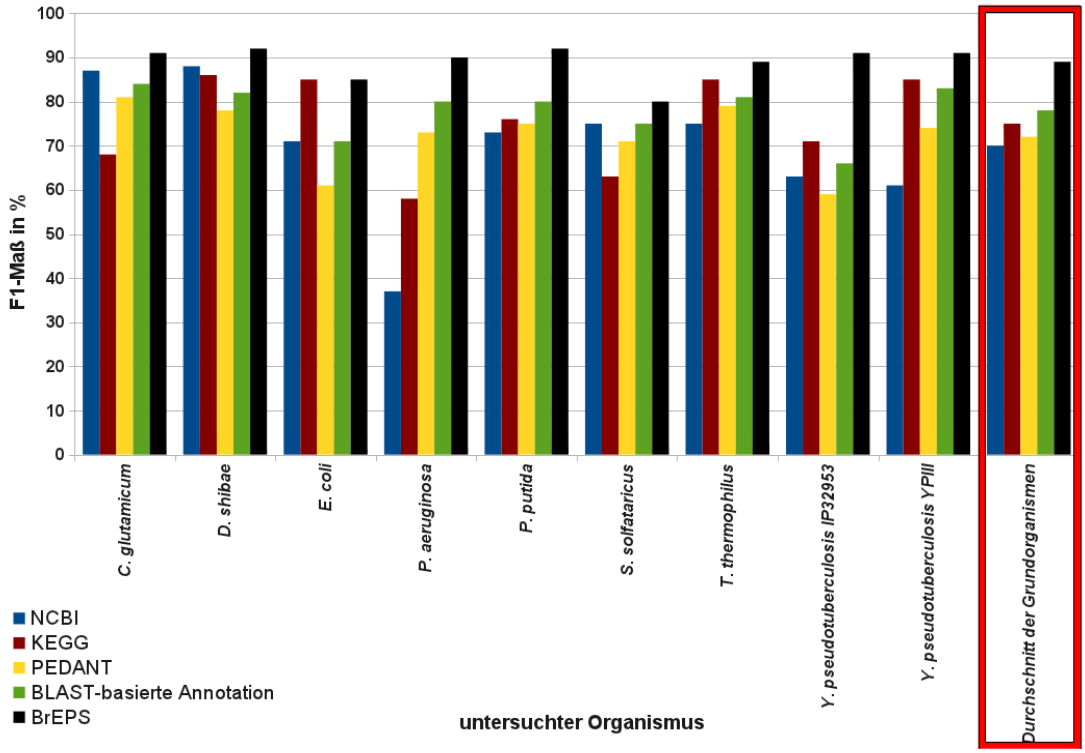


Abbildung 6 - F1-Maß der einzelnen Annotationsquellen der Referenzorganismen und der Durchschnittswert dieser neun Organismen

Wie unter Punkt 2.12 auf Seite 34 beschrieben, wurden den einzelnen Annotationsquellen basiert auf dem Vergleich der Annotationen mit Swiss-Prot ein

Relevanzwert zugeordnet. Entscheidungsgrundlage waren hier die F1-Maße der einzelnen Datenbanken.

Einem F1-Maß von genau 100 % wurde eine Einzelrelevanz von 13 zugeordnet. Für ein F1-Maß zwischen 95 % und 100 % wurde eine Relevanz von 12 festgelegt. Für alle anderen Werte fällt die Einzelrelevanz um 1 für jeden Abstieg des F1-Maßes um 5 %. Somit bekommen F1-Maße unter 40 % eine Einzelrelevanz von 0 zugeordnet.

Die Werte wurden so gewählt, da der BLAST-basierten Genomannotation aus programmtechnischen Gründen eine Einzelrelevanz von 8 (oder einem Vielfachen davon) zugeordnet werden sollte. Entsprechend wurde das F1-Maß der BLAST-basierten Genomannotation ermittelt, der Wert der Relevanz auf 8 festgelegt und die Relevanzen der anderen Quellen entsprechend verteilt.

Die Ergebnisse der Berechnung der F1-Maße sind in Abbildung 6 zu sehen. Die durchschnittlichen F1-Werte, die für die Ermittlung der Standardwerte herangezogen wurden, sind in durch eine rote Umrandung hervorgehoben.

Auffällig sind die F1-Werte für P. *aeruginosa* für die Datenbanken NCBI und KEGG. Diese fallen deutlich niedriger aus als bei den anderen Organismen. Erklären lässt sich dies dadurch, dass in diesen beiden Datenbanken die Anzahl der vorhandenen Annotationen sehr gering ausfiel. Zum Zeitpunkt der Berechnung der Daten lag in beiden Datenbanken nur eine Genomannotation von Mai 2010 vor. Diese wurde inzwischen erneuert und bietet nun auch mehr Informationen. Da die Anzahl der Annotationen gering war, fiel auch die Überschneidung mit Swiss-Prot nicht sehr hoch aus und dies führte zu den niedrigen *Precision*- und *Recall*- Werten. Nach einem Update der Daten mit den neuen Genomannotationsdateien der beiden Datenbanken wird dies wahrscheinlich anders aussehen.

Gemäß den Ergebnissen des Vergleichs der Quellen mit Swiss-Prot wurde KEGG und PEDANT jeweils eine Einzelrelevanz von 7 zugeordnet, NCBI die Einzelrelevanz von 6.

Da für die BLAST-basierte Genomannotation ein klares Gütemerkmal, der E-value, vorhanden ist, wird hier die Einzelrelevanz entsprechend diesem Gütemerkmal

zugeteilt. Es wurde ein maximaler Score von 8 festgelegt. Der Einzelscore für die BLAST-basierte Genomannotation setzt sich aus zwei Teilen zusammen, der Auswertung des besten E-values, mit dem die Annotation in der gesamten UniProtKB gefunden wurde, und der Auswertung des E-values, mit dem die Annotation in Swiss-Prot gefunden wurde. Für die Auswertung der E-values wurden vier Gruppen angelegt mit entsprechenden Relevanzwerten. Annotationen mit einem E-value größer als 10^{-40} wird ein Wert von 1 zugeordnet, Annotationen mit einem E-value zwischen 10^{-40} und 10^{-80} ein Wert von 2, solchen mit einem E-value zwischen 10^{-80} und 10^{-120} ein Wert von 3 und Annotationen mit einem E-value kleiner 10^{-120} der maximale Wert von 4. Aus der Summe der Gruppenzuweisungen der beiden E-values ergibt sich die Gesamt-Einzelrelevanz, die der BLAST-basierten Genomannotation zugeordnet wird.

Auch für die BrEPS-Ergebnisse ist ein Gütemerkmal, die *Correctness* vorhanden[33]. Es wurde basierend auf dem F1-Maß ein maximaler Wert von 10 bestimmt. Abhängig von *Correctness* und der Anzahl von Enzymen innerhalb des BrEPS-Knoten wird die Annotation in eine von 6 Gruppen eingeteilt und entsprechend eine Einzelrelevanz zwischen 1 und 10 zugewiesen.

Manuell überprüfte Annotationen, wie denen aus UniProt und BRENDA, wurde eine Einzelrelevanz von 50 zugewiesen. Somit ist der Relevanzwert für die manuell überprüften Informationen stets deutlich höher als die Summe der Einzelrelevanzen aller anderen Quellen. Dadurch ist sichergestellt, dass Informationen dieser Güte nicht durch die anderen Quellen überstimmt werden.

Annotationen, die aus AMENDA erhalten werden, wurde eine Einzelrelevanz von 25 zugewiesen. Die Daten haben eine hohe Zuverlässigkeit, sind aber nicht ganz so sicher wie manuell überprüfte Daten. Es handelt sich zwar um Daten aus der Primärliteratur, allerdings werden sie mittels eines *Textmining*-Programms extrahiert und nicht manuell überprüft.

Die Summe aller dieser Einzelrelevanzen ergibt die Gesamtrelevanz einer Annotation. Sie steht für die Qualität der Annotation und ermöglicht dem Nutzer, zwischen mehreren Kandidaten abzuwägen.

3.1.2 Annotationsergebnisse aus C. glutamicum

Um den Umfang der möglichen Ergebnisse des Programms vorzustellen, werden im Folgenden einige Annotationsergebnisse für *C. glutamicum* betrachtet. Die verschiedenen Beispiele sind in Tabelle 6 aufgelistet.

Tabelle 6 - Beispiele für Annotationsergebnisse für C. glutamicum aus dem Programm EnzymeDetector

Nr.	GI	Locus-Tag	EC	E-value	Gesamtrelevanz	Quellen
1	62390419	cg1734	4.99.1.1	0.0	92	NCBI, KEGG, PEDANT, BLAST-basiert, BrEPS, UniProt
2	62390772	cg2116	2.7.1.56	0.0	11	BLAST-basiert, PEDANT
			2.7.1.11	1E-28	7	PEDANT
3	62389677	cg0898	1.1.1.262	7E-103	3	BLAST-basiert
			5.1.3.9	7E-93	3	BLAST-basiert
4	62389751	cg0984	2.1.2.3	0.0	90	NCBI, KEGG, PEDANT, BLAST-basiert, BrEPS, UniProt
			3.5.4.10	0.0	86	NCBI, KEGG, PEDANT, BLAST-basiert, BrEPS, UniProt

In Tabelle 6 ist unter **Beispiel 1** ein eindeutiges Ergebnis aufgelistet. Es gibt nur einen Annotationskandidaten für das Gen. Die vorgeschlagene EC-Nummer wird nicht nur in der BLAST-Suche mit einem perfekten E-value gefunden, sondern auch von NCBI, KEGG, PEDANT, BrEPS und UniProt bestätigt. Dies ist ein Beispiel für eine Annotation, an der keine Zweifel bestehen und bei der eine manuelle Untersuchung des Ergebnisses nicht nötig ist.

Im **Beispiel 2** werden zwei Kandidaten gefunden. Beide Annotationen werden von PEDANT genannt, allerdings wird nur eine der Annotationen durch die BLAST-basierte Genomannotation unterstützt. Dies lässt sich klar an den E-values erklären, da der eine Kandidat mit einem E-value von 0.0 gefunden wird und der andere im Vergleich dazu nur mit einem E-value von 10^{-28}. Dies ist ein Beispiel für den Fall, dass die BLAST-

basierte Genomannotation eine Entscheidungshilfe darstellen kann, wenn eine Datenbank zwei Annotationen angibt.

Beispiel 3 zeigt einen Fall, bei dem durch die BLAST-basierte Genomannotation zwei EC-Nummern gefunden werden. Beide EC-Nummern werden mit sehr ähnlichem E-value gefunden. Da diese Annotationen ausschließlich durch die BLAST-basierte Genomannotation gefunden werden, erhalten beide die gleiche Gesamtrelevanz. In diesem Fall ist es nicht möglich, automatisch zu entscheiden, welcher der beiden Kandidaten bevorzugt werden soll. Auf Grund dessen wäre hier die Empfehlung, beide zu übernehmen. Eine Alternative wäre die manuelle Untersuchung dieses Genabschnitts und eine anschließende manuelle Entscheidung für einen der beiden Kandidaten. Ein solcher Genabschnitt eignet sich auch für die weitere Untersuchung im Labor, da nun gezielt nach den zwei Kandidaten untersucht werden kann. Dies bietet sich an, falls einer der gefundenen Kandidaten eine entscheidende Rolle im Stoffwechsel des Organismus einnimmt.

In **Beispiel 4** gibt es ebenfalls zwei Kandidaten mit sehr ähnlicher Gesamtrelevanz. Allerdings werden diese beiden Kandidaten in mehreren Quellen gefunden, und somit ist ihre Gesamtrelevanz jeweils sehr hoch. Dies spricht dafür, dass das Genprodukt tatsächlich beide Enzymfunktionen besitzt. Deshalb sollte auch keine Entweder-oder-Entscheidung getroffen werden, sondern es sollten beide Enzyme gleichwertig übernommen werden.

Die Beispiele zeigen, dass in vielen Fällen eine zuverlässige automatische Entscheidung möglich ist, es aber auch einige Fälle gibt, bei denen dies schwierig ist. Bei manchen dieser unklaren Annotationen kann durch den Zusammenschluss der Informationsquellen eine klare Entscheidung herbeigeführt werden, bei manchen Genen ist eine Entscheidung aber trotzdem nicht möglich, ohne eine manuelle weiterführende Untersuchung des Gens vorzunehmen. Hier hilft der EnzymeDetector, die plausiblen Kandidaten für eine Annotation einzugrenzen.

3.1.3 Anteil von Enzymgenen am Genom

Für die Ermittlung des Anteils an vorhergesagten Genen, die für Enzyme kodieren, wurden nur Ergebnisse einbezogen, die eine minimale Gesamtrelevanz von 7 erreichten. In Tabelle 7 sind die Werte der 9 Referenzorganismen und der Durchschnittswert aller untersuchten Organismen aufgeführt.

Tabelle 7 - Anteil an Genen des Gesamtgenoms, für die eine Enzymfunktion durch das Programm EnzymeDetector vorhergesagt wurde

Organismus	**Anteil der Gene mit durch EnzymeDetector vorhergesagter Enzymfunktion**
C. glutamicum	29 %
D. shibae	36 %
E. coli	47 %
P. aeruginosa	27 %
P. putida	26 %
S. solfataricus	25 %
T. thermophilus	27 %
Y. pseudotuberculosis IP32953	26 %
Y. pseudotuberculosis YPIII	31 %
Durchschnitt aller Organismen	46 %

Durchschnittlich wurde für 30 % der Gene der neun Referenzorganismen eine Enzymfunktion vorhergesagt. Dieser Wert passt zum allgemein akzeptierten Anteil an Enzymgenen in prokaryotischen Genomen. Als Referenzwert fungiert der Enzymgehalt von *Escherichia coli* laut Swiss-Prot, da dieser Organismus einer der bestuntersuchten ist. Laut UniProt kodieren 35 % der Gene von *E. coli* für Enzyme.

Der durchschnittliche Wert, gemittelt über alle 90 Organismen, liegt mit 46 % deutlich höher. Diese Abweichung lässt sich dadurch erklären, dass unter den 90 Organismen auch bisher schlecht untersuchte vorkommen. Für diese sind allgemein noch nicht so viele Gene bekannt. Dies ist in Tabelle 8 verdeutlicht. Für *E. coli* sind für den

untersuchten Abschnitt von 1 Millionen Basenpaaren 883 Gene bekannt. Für die 89 anderen untersuchten Organismen ist die Anzahl der Gene für diese 1 Millionen Basenpaare sehr unterschiedlich verteilt. Immerhin fast 1/3 der Organismen hat eine Genanzahl von unter 400 Genen für den Abschnitt von 1 Millionen Basenpaaren.

Tabelle 8 - Anzahl der Gene pro 1 Millionen Basenpaare (BP). Der Wert für E. coli ist als Referenz angegeben. Die 89 anderen Organismen sind in 100er Gruppen eingeordnet.

E. coli Genanzahl / 1 Mio. BP	**883**
Genanzahl / 1 Mio. BP	**Anteil der 89 Organismen**
0 – 99	0 %
100 – 199	6 %
200 – 299	15 %
300 – 399	6 %
400 – 499	2 %
500 – 599	2 %
600 – 699	1 %
700 – 799	2 %
800 – 899	26 %
900 – 999	31 %
> 1000	8 %

Hinzu kommt, dass für die schlechter untersuchten Organismen meist die Gene bekannt sind, die eine enzymatische Funktion tragen. Somit liegt der Anteil der Gene mit zugeordneter Funktion für schlechter untersuchte Organismen höher als für sehr gut untersuchte. Diese Tatsache sorgt dafür, dass der durchschnittliche Wert des Anteils an Enzymgenen für alle 90 Organismen höher liegt als der Erwartungswert von 35 %.

3.1.4 Verteilung der Enzyme auf die Hauptklassen

Tabelle 9 führt die Anzahl der ermittelten Enzyme für die einzelnen Genome und die Verteilung dieser Enzyme auf die 6 Enzymhauptklassen auf.

Auch wenn die absolute Anzahl an gefundenen Enzymen zwischen den Organismen stark variiert, so ist ihre Zusammensetzung bezüglich der Hauptklassen sehr ähnlich verteilt. Für die Hauptklasse 4 liegt der Anteil an Enzymen für alle Organismen

zwischen 9 und 10 % und für Hauptklasse 5 zwischen 4 und 6 %. Knapp ein Drittel aller Enzyme verteilt sich bei allen Organismen auf die ersten drei Hauptklassen.

Tabelle 9 - Anzahl an Enzymen, die im Organismus gefunden wurden, und ihre Verteilung auf die 6 Hauptklassen

Organismus	**Anzahl Enzyme**	**1**	**2**	**3**	**4**	**5**	**6**
C. glutamicum	1389	28 %	28 %	24 %	9 %	5 %	7 %
D. shibae	1496	29 %	26 %	24 %	10 %	5 %	6 %
E. coli	3023	25 %	28 %	29 %	9 %	5 %	4 %
P. aeruginosa	2009	30 %	25 %	26 %	9 %	4 %	5 %
P. putida	1845	32 %	24 %	24 %	9 %	5 %	6 %
S. solfataricus	1233	24 %	29 %	25 %	9 %	5 %	8 %
T. thermophilus	1380	26 %	29 %	24 %	9 %	4 %	7 %
Y. pseudotuberculosis IP32953	1616	27 %	27 %	24 %	10 %	6 %	6 %
Y. pseudotuberculosis YPIII	1553	26 %	27 %	25 %	10 %	6 %	7 %
Durchschnitt	1722	26 %	28 %	26 %	10 %	5 %	7 %

Die gefundene Verteilung in den Organismen passt zu der allgemeinen Verteilung der Enzyme auf die Hauptklassen. 27% aller Enzyme gehören zu Hauptklasse 1, 28 % zu Hauptklasse 2, 29 % zu Hauptklasse 3. In Hauptklasse 4 finden sich 10 % aller Enzyme und in den Hauptklassen 5 und 6 jeweils 4 bzw. 3 % aller Enzyme. Die Durchschnittswerte aus der Tabelle decken diese Werte sehr gut ab.

Es ist interessant zu sehen, dass, auch wenn die verschiedenen Organismen sehr unterschiedliche Anforderungen an ihren Enzymhaushalt haben und teilweise sehr spezialisiert sind, der Anteil der Vertreter einer bestimmten Reaktionsklasse von Enzymen gleichmäßig verteilt ist.

Die Anteile der Enzymklassen, die sowohl unter den Organismen sehr stimmig sind, als auch mit der Verteilung der Enzyme auf die Hauptklassen übereinstimmen, sprechen für die Stabilität der EnzymeDetector Ergebnisse. Würden durch die BLAST-basierte

Genomannotation unsinnige Enzymfunktionen zur Gesamtannotation hinzugefügt werden, würde sich das in der Verteilung der Ergebnisse auf die Hauptklassen niederschlagen.

3.1.5 Informationsgehalt der unterschiedlichen Annotationsquellen

Betrachtet man alle Gene, zu denen durch das Programm EnzymeDetector in einer der Quellen eine Annotation gefunden wurde, zeigen sich deutliche Unterschiede im Informationsgehalt dieser Quellen. In Abbildung 7 ist der Anteil der Annotationsquellen an Genen mit annotierter Enzymfunktion aufgeführt. Diese Werte stellen den Durchschnitt über alle 90 untersuchten Organismen dar.

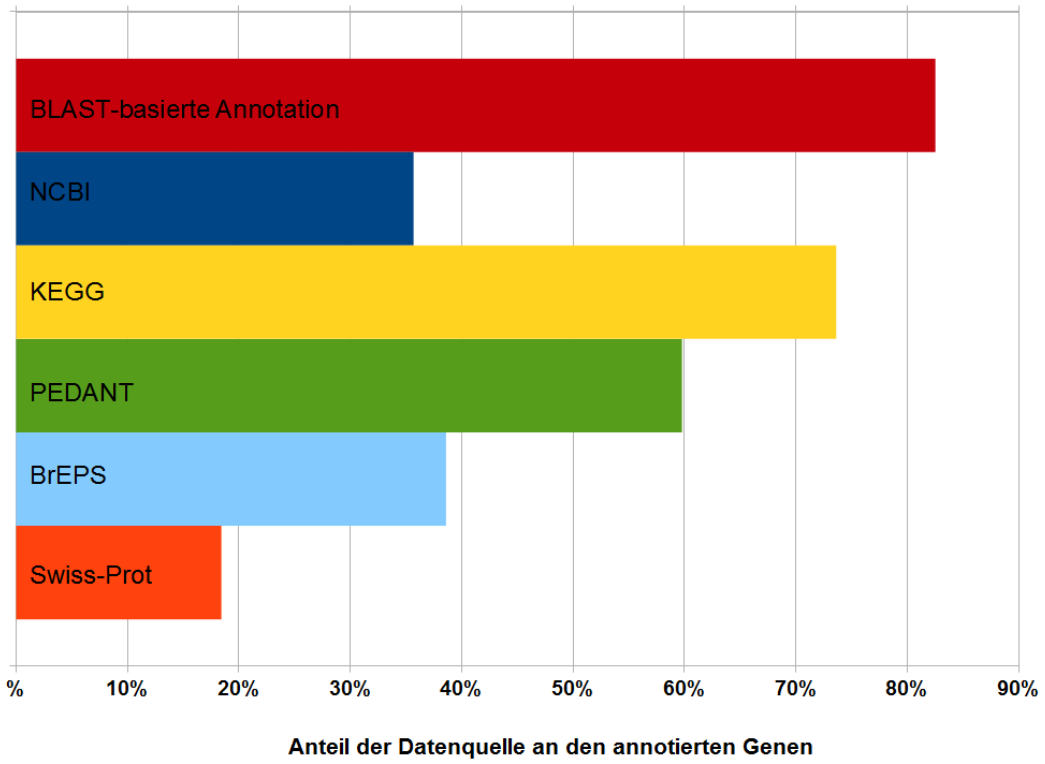


Abbildung 7 - Anteil der verschiedenen Annotationsquellen an den annotierten Genen der Organismen

Ein großer Teil der Annotationen wird durch die BLAST-basierte Genomannotation geliefert. Viele werden sogar ausschließlich durch die BLAST-Suche gefunden, was den

Wert dieser Annotationen deutlich erhöht (siehe Punkt 3.1.6 Seite 57). Gewertet wurden für diese Analyse nur Annotationen, die höchstens einen E-value von 10^{-80} hatten.

Von den anderen Quellen sind in PEDANT und KEGG die meisten Gene annotiert. In NCBI und BrEPS liegen deutlich weniger annotierte Gene vor. Dies deckt sich mit der Erfahrung im Umgang mit diesen Datenbanken.

Für durchschnittlich 18 % aller Gene gibt es eine Swiss-Prot-Annotation. Diesen Wert sollte man bei allen Analysen, bei denen die Swiss-Prot-Genomannotation als Wahrheitsmaß dient, berücksichtigen.

Die Werte zeigen deutlich, dass das Arbeiten mit nur einer der Quellen deutliche Verluste im Informationsgehalt mit sich bringen muss. Ohne die Integration von Daten mehrerer Quellen würde der angenommene Anteil von 35 % der Genen, die für Enzyme kodieren, niemals erreicht werden.

3.1.6 *Übereinstimmung der Annotationen zwischen den Annotationsquellen*

Nicht nur der Informationsgehalt der Datenbanken unterscheidet sich deutlich. Oft stimmen die Datenbanken auch für die Gene, für die eine Annotation in mehreren Quellen vorhanden ist, nicht überein.

Um dies zu verdeutlichen, wurden die Annotationen der drei Hauptannotations-Datenbanken NCBI, KEGG und PEDANT und die BLAST-basierte Genomannotation miteinander verglichen. Es wurden nur Organismen in den Vergleich einbezogen, für die es eine Genomannotations-Datei in allen drei Datenbanken gab, damit die Daten gleichwertig untereinander verglichen werden konnten. Dies war für 45 der 90 Organismen der Fall.

Die durchschnittliche Übereinstimmung der Datenbanken ist in Abbildung 8 zu sehen. Nur in 23,6 % aller Annotationen war die Annotation der drei Datenbanken identisch. In weiteren 26 % stimmten zwei der drei Quellen überein und in 31 % kam eine Annotation nur in einer der drei Quellen überhaupt vor. Die Schnittmenge zwischen allen vier Annotationsquellen beträgt 23,2 %.

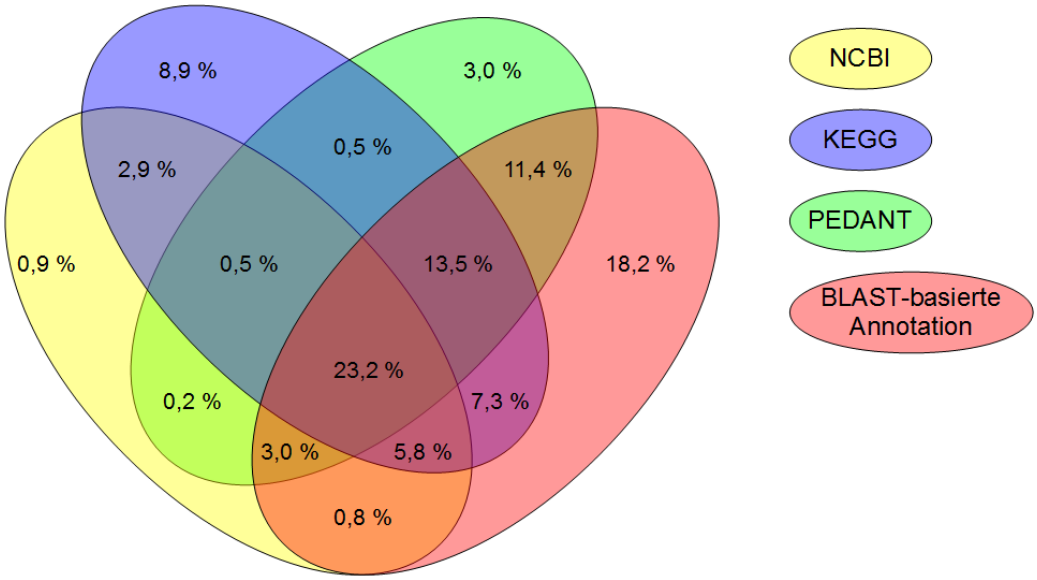


Abbildung 8 - Übereinstimmung der Annotationen der drei Annotationsdatenbanken NCBI, KEGG und PEDANT und der BLAST-basierten Genomannotation

Die Unterschiede in den Datenbanken unterstreichen die Wichtigkeit eines Programmes wie des EnzymeDetectors. Ohne solche Programme hat der Nutzer keine Möglichkeit, zwischen Annotationen hoher und niedriger Qualität zu unterscheiden. Auch ist es unmöglich abzuschätzen, welche Gründe für die unterschiedlichen Annotationen vorliegen.

Neben der Schwierigkeit zwischen Annotationen unterschiedlicher Qualität bzw. Richtigkeit zu unterscheiden, zeigen die Zahlen deutlich, dass erst mit den kombinierten Ergebnissen aller Quellen eine zufriedenstellende Informationsabdeckung erreicht werden kann. Denn jede der Quellen liefert einen Anteil an Annotationen, die alleine durch sie beigesteuert werden.

18 % der Annotationen wurden nur durch die BLAST-basierte Genomannotation gefunden. Diese Annotationen waren in keiner der drei Datenbanken zu finden. Bei diesen Daten handelt es sich aber nicht um Zufallstreffer auf Grund von leichten Sequenzähnlichkeiten mit hohen E-values. Für die BLAST-basierte Genomannotation wurden in diese Analyse nur Treffer einbezogen, die höchstens einen E-value von 10^{-80} hatten, also eindeutig signifikante Treffer sind.

Diese Treffer lassen sich dadurch erklären, dass für die BLAST-basierte Genomannotation die zu diesem Zeitpunkt aktuellsten verfügbaren Sequenzen verwendet wurden. Die Annotationen, die sich in großen Datenbanken wie NCBI, KEGG und PEDANT finden, werden zwar regelmäßig aktualisiert, allerdings finden sich teilweise Annotationsdaten, die über ein Jahr alt sind. Viele der Sequenzen, die sich aktuell in den Datenbanken befinden, gab es zu dem Zeitpunkt noch nicht, und somit konnten sie nicht für die Funktionsvorhersage in Betracht gezogen werden. Auf Grund der Fülle der Daten ist es den großen Datenbanken nicht möglich, kürzere Aktualisierungsintervalle für alle bekannten Organismen zu realisieren.

Weiterhin führen die Unterschiede in den Entscheidungskriterien für eine Annotation zu diesen Unterschieden.

Doch auch die anderen Annotationsquellen lieferten jeweils Daten, die von keiner der anderen Quelle geliefert wurden. KEGG bietet hier mit 9 % den größten Anteil an Zusatzdaten neben der BLAST-basierten Genomannotation. PEDANT lieferte 3 % der Annotationen ausschließlich und NCBI liefert nur 1 % an zusätzlichen Informationen. Dies lässt sich mit der unterschiedlichen Art des Annotierens der Quellen und den unterschiedlichen Aktualisierungszyklen begründen. Auch spielt es eine Rolle, ob nur hand-annotierte oder auch computergestützte Annotationen in die Datenbank einbezogen werden und in welchem Umfang dies geschieht.

3.1.7 Vergleich der Annotationen auf Ebene von EC-Nummern

Für den Teil der Gene, die sowohl in der BLAST-basierten Genomannotation als auch in einer der anderen Annotationsquellen annotiert waren, wurden die Annotationen auf Ebene der EC-Nummern genauer miteinander verglichen. Diese Analyse wurde durchgeführt, um zu ermitteln, ob, und wenn ja, inwieweit die BLAST-basierte Genomannotation mit den Genomannotationen der anderen Quellen übereinstimmte.

Die Übereinstimmung zwischen den Annotationen wurde in vier Gruppen kategorisiert. Diese sind in Tabelle 10 mit Beispielen veranschaulicht.

Tabelle 10 - Übereinstimmungskategorien beim Vergleich der Annotationen zweier Annotationsquellen

Kategorie	Annotation Quelle 1	Annotation Quelle 2
identisch	1.1.1.1 1.1.1.1, 2.1.1.2	1.1.1.1 1.1.1.1, 2.1.1.2
teilweise identisch ohne Widerspruch	1.1.1.1	1.1.1.1, 2.1.1.2
teilweise identisch mit Widerspruch	1.1.1.1, 2.1.1.2	1.1.1.1, 2.1.2.3
Widerspruch	1.1.1.1	2.1.1.1

Die Fälle mit einer identischen oder einer widersprüchlichen Annotation für ein Gen sind klar abgegrenzt. Die Fälle von Annotationen mit einer teilweisen Übereinstimmung wurden in zwei Kategorien unterteilt. Diese Unterscheidung wurde vorgenommen, da es ein größerer Widerspruch ist, wenn auch unterschiedliche Annotationen zu finden sind, als wenn es sich nur um das Fehlen einer Annotation handelt.

Die Ergebnisse dieser Kategorisierung sind in Abbildung 9 zu sehen. Der Vergleich wurde für die Quellen NCBI, KEGG, PEDANT, BrEPS und Swiss-Prot durchgeführt. Die Quellen BRENDA und AMENDA wurden nicht berücksichtigt, da die zugehörigen Annotationen zum größten Teil sequenzunspezifisch vorliegen und sich somit nicht vergleichen lassen.

Die Ergebnisse wurden für den Vergleich der BLAST-basierten Genomannotation mit jeder der Annotationsquellen in die vier beschriebenen Kategorien eingeteilt. Hierfür wurden die Daten über alle Organismen gemittelt.

Für einen Großteil der Gene, die sowohl in der BLAST-basierten Genomannotation als auch in der jeweiligen Datenbank annotiert waren, finden sich identische Annotationen in den verglichenen Quellen. So waren zum Beispiel die Annotationen aus PEDANT und der BLAST-basierten Genomannotation in 74 % der Fälle identisch. Für NCBI waren sie es sogar in 75 % der Fälle. Für mindestens 10 % der Fälle wurde beim Vergleich mit der BLAST-basierten Genomannotation zumindest keine widersprüchliche Information gefunden. Beim Vergleich mit BrEPS traf dies sogar auf 36 % zu. Es überrascht nicht, da sowohl die BLAST-basierte Genomannotation als auch

die BrEPS Vorhersage mit Hilfe der Swiss-Prot-Datenbank hergestellt wurde. Allerdings werden für die BLAST-basierte Genomannotation auch die TrEMBL-Daten einbezogen.

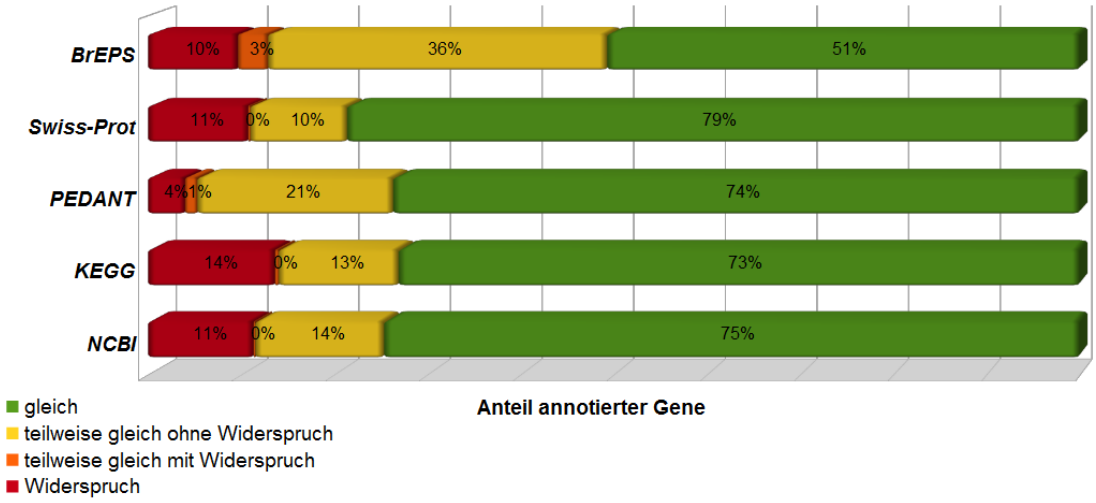


Abbildung 9 - Grad der Übereinstimmung der Annotationen der Gene, die sowohl in der BLAST-basierten Genomannotation als auch der jeweiligen Quelle annotiert waren

Für jeweils weniger als 14 % der Fälle fand sich ein partieller oder kompletter Widerspruch zwischen der Annotation aus der BLAST-basierten Genomannotation und der Annotation durch die Datenbank, wobei der Vergleich mit KEGG den größten Widerspruch zeigt. KEGG ist allerdings neben der BLAST-Suche auch die Quelle mit den meisten Annotationen. Es ist also nicht verwunderlich, dass es hier häufiger zu Abweichungen kommt.

Diese Werte zeigen, dass die BLAST-basierte Genomannotation einen Mehrwert an Annotationen liefert, und in den Fällen, wo Annotationen schon in den Datenbanken vorhanden sind, diese eher bekräftigt, als ihnen zu widersprechen.

Für den Fall, dass es in den Datenbanken mehrere Enzymfunktionen zu einem Gen gibt, kann die BLAST-basierte Genomannotation eine große Hilfe darstellen bei der Abschätzung, welche der Enzymfunktionen wirklich vorhanden sind bzw. ob dem Gen eventuell mehrere Enzymfunktionen zugeordnet werden müssen.

Die größte Übereinstimmung hat die BLAST-basierte Genomannotation mit Swiss-Prot. Dies ist darauf zurückzuführen, dass gegen die vollständige UniProt Datenbank

geblastet wird, die auch Swiss-Prot beinhaltet. Die Übereinstimmung beträgt nicht 100 %. Einerseits sind in Swiss-Prot unvollständige EC-Nummern annotiert, diese werden aber nicht in die BLAST-basierte Genomannotation übernommen. Andererseits werden zusätzlich zu den in Swiss-Prot annotierten Enzymen durch die BLAST-Suche noch weitere Enzyme gefunden und dem Gen als Kandidat zugeordnet. Somit kann ein Gen in eine der beiden mittleren Gruppen fallen, auch wenn es eine Übereinstimmung mit Swiss-Prot gibt.

Die Werte zeigen, dass die BLAST-basierte Genomannotation mit den Datenbanken NCBI, KEGG und PEDANT sehr gut überstimmt. Dies legt nahe, dass die nicht veröffentlichten Entscheidungskriterien für die Auswahl der Annotationen der Datenbanken denen ähneln, die für die BLAST-basierte Genomannotation gewählt wurden.

3.1.8 E-value Bereiche der BLAST-basierten Genomannotation

Die E-values der Annotationen, die ausschließlich durch die BLAST-basierte Genomannotation gefunden wurden, wurden in Güteklassen eingeteilt. Abbildung 10 gibt einen Überblick über diese Einteilung.

Durchschnittlich 25 % dieser Annotationen haben einen niedrigen E-value zwischen 10^{-50} und 10^{-120}. 29 % haben sogar einen sehr niedrigen E-value von kleiner 10^{-120}.

Die 20 % der Annotationen, die im E-value-Bereich zwischen 10^{-20} und 10^{-50} liegen, stellen gute Kandidaten dar, wenn es darum geht, eine fehlende Enzymaktivität für die Erstellung eines metabolischen Modells einem Gen des Organismus zuzuordnen. 27 % der Annotationen hatten E-values größer als 10^{-20}.

Der hohe Anteil an BLAST-Treffern mit niedrigen E-values ($<10^{-20}$) resultiert daraus, dass die BLAST-basierte Genomannotation für jedes Gen eine Suche durchführt und das beste Ergebnis aus dieser Suche als Kandidaten übernimmt. Gibt es nur Ergebnisse mit schlechten E-values, so werden diese als mögliche Kandidaten übernommen. Allerdings erhalten diese Treffer in der Auswertung einen niedrigen Relevanzwert, da dieser ja für die BLAST-basierte Genomannotation E-value-bezogen ist. Somit werden diese

schlechten Treffer automatisch aus der Analyse ausgeschlossen, wenn ein entsprechender Grenzwert gewählt wird. Hierfür reicht schon die Wahl des niedrigsten empfohlenen Grenzwertes von 7.

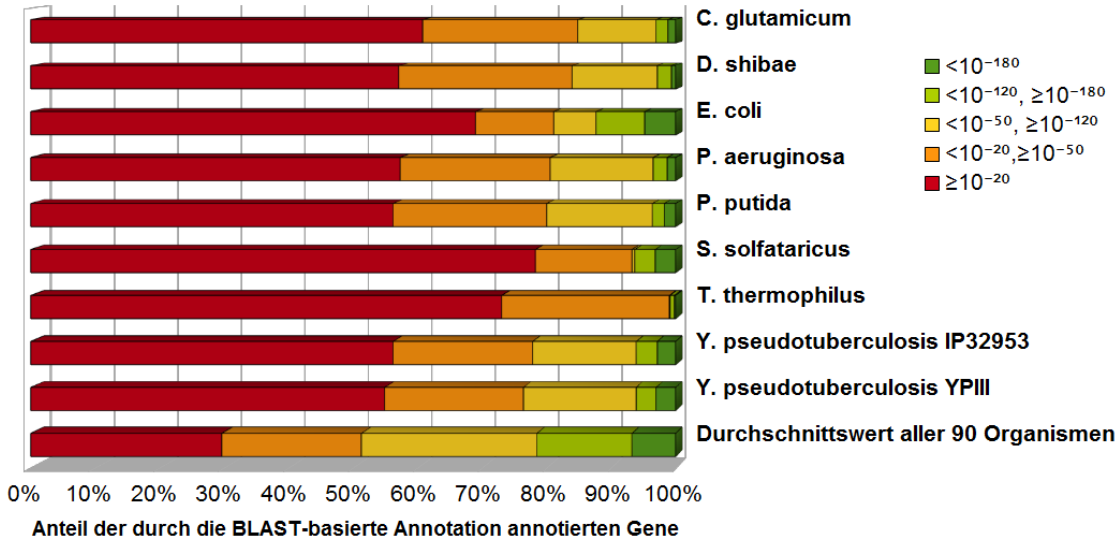


Abbildung 10 - E-value-Güteklassen der Annotationen aus der BLAST-basierten Genomannotation (Abbildung groß unter Anhang 3)

Die Werte zeigen aber auch, dass ein großer Teil der gefundenen zusätzlichen Annotationen in niedrigen E-value-Bereichen liegt. Das heißt, dass eine Vielzahl der Gene, die durch die BLAST-basierte Genomannotation neu annotiert werden konnten, wertvolle und signifikante Informationen zum Informations-Pool hinzufügen.

Vergleicht man die Werte der neun Referenzorganismen, so wird deutlich, dass die zusätzlichen Annotationen für das hyperthermophile Archaeon *S. solfataricus* eine niedrigere Qualität haben als die der analysierten Bakterien. Dies entspricht der Erwartung, da es wenige verlässliche Enzymsequenzen für Archaeen gibt und diese einen hochspezialisierten Metabolismus haben.

Auch fällt auf, dass die Ergebnisse für *E. coli* häufiger in den sehr niedrigen E-value-Bereichen liegen und somit im Vergleich zu den anderen Organismen eine deutlich höhere Qualität aufweisen. Da *E. coli* einer der bestuntersuchten Organismen überhaupt ist und somit sehr viele Sequenzinformationen zu den unterschiedlichen Stämmen in UniProt vorliegen, die bei einer BLAST-Auswertung zum Tragen kommen, ist dies nicht überraschend. Es ist aber auch zu sehen, dass der Anteil der Annotationen, die im

schlechten E-value-Bereich liegen, bei *E. coli* etwas erhöht ist im Gegensatz zu den anderen untersuchten Bakterien. Es werden also für viele Gene ausschließlich Treffer mit schlechten E-values gefunden. Man kann bei diesen Genen davon ausgehen, dass es sich nicht um Gene mit Enzymfunktionen handelt, sondern um Gene mit Ähnlichkeiten zu Sequenzen mit Enzymfunktion. Diese Ergebnisse können ignoriert werden und sollten in den Ergebnispool nicht einbezogen werden.

3.1.9 Verteilung der Gesamtrelevanzen der Ergebnisse

Um die Ergebnisse des EnzymeDetectors näher zu beleuchten, wurden diese nach ihren Gesamtrelevanzen gruppiert. Die Einteilung in Gruppen erfolgte nach dem Grenzwertschema, was für die Webseite ermittelt wurde (siehe Punkt 3.3.2 auf Seite 74).

Annotationen mit einer Gesamtrelevanz kleiner als sieben, d.h. kleiner als der minimale Grenzwert, sind in Abbildung 11 in Rot zu sehen. Durchschnittlich gehören 32 % aller Annotationen in diese Gruppe.

Qualitativ gute Annotationen mit einer Gesamtrelevanz zwischen 7 und 25 sind in der Abbildung in Orange dargestellt. Zu dieser Gruppe gehören durchschnittlich 50 % der Annotationen.

Annotationen mit einer sehr guten Sicherheit liegen in einem Gesamtrelevanz-Bereich zwischen 26 und 38. Diese Ergebnisse, dargestellt in Gelb, haben einen perfekten *Recall* und eine *Precision* von über 98 %. Es finden sich 4 % der Annotationen in diesem Bereich.

Annotationen der besten Gruppe mit einer Gesamtrelevanz größer als 38, dargestellt in Grün, haben, je höher die Gesamtrelevanz geht, zwar nicht mehr einen so hohen *Recall*, aber die *Precision* liegt bei 100 %. Immerhin noch 17 % der Annotationen liegen in diesem Relevanzbereich.

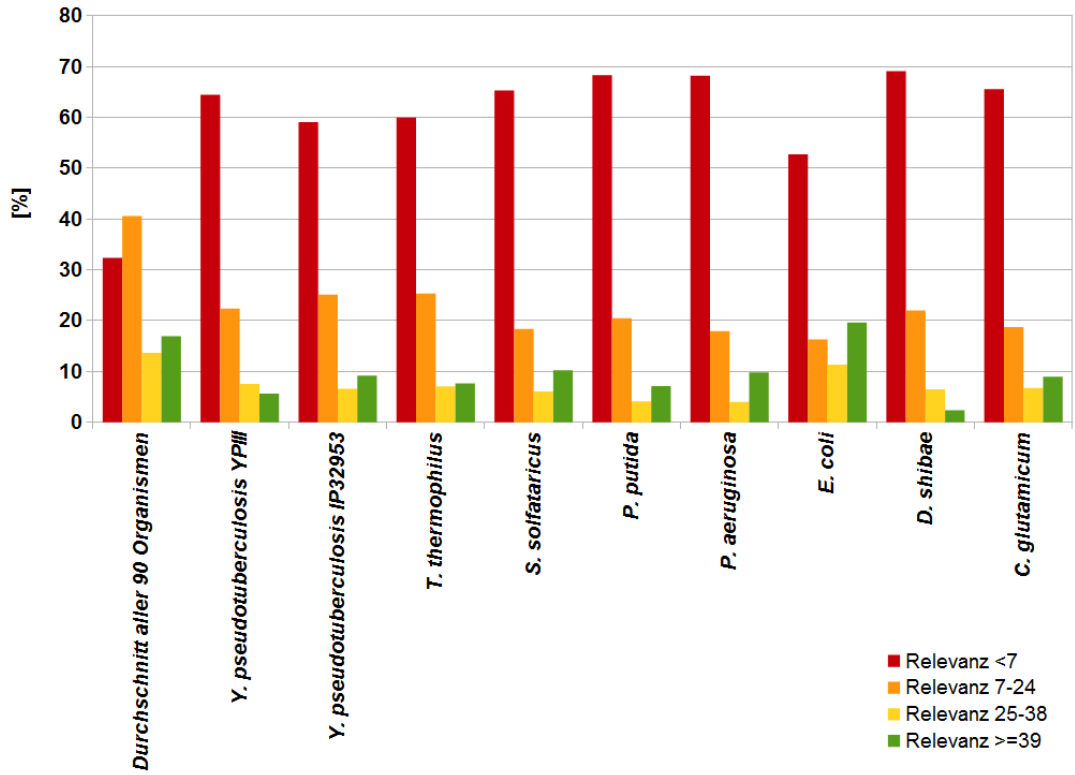


Abbildung 11 - Relevanzgruppen der EnzymeDetector Ergebnisse der neun Referenzorganismen und der Durchschnitt dieser Werte über alle 90 Organismen (Abbildung groß unter Anhang 4)

Der Anteil der Annotationen, die zur niedrigsten Relevanzgruppe gehören, scheint relativ hoch. Diese Ergebnisse resultieren aber zum größten Teil aus BLAST-Ergebnissen mit hohen oder mittleren E-values. Diese werden von der BLAST-basierten Genomannotation als Kandidaten übernommen, wenn sich keine Treffer mit kleineren E-values finden. Es handelt sich bei diesen Annotationskandidaten allerdings eher um Datenmüll als wirkliche Kandidaten für eine sinnvolle Annotation. Diese Tatsache wird aber mit der Wahl eines geeigneten Grenzwertes relativiert. Selbst wenn der niedrigste Grenzwert gewählt wird, der im Grenzwertschema ermittelt wurde, werden diese schlechten Kandidaten nicht in das Endergebnis übernommen.

Auch Annotationen, die in nur einer der anderen Quellen vorkommen, können zu diesem Relevanzbereich gehören. Dies trifft aber nur für NCBI zum, da dieser Datenbank auf Grund ihrer Bewertung gegen Swiss-Prot eine Einzelrelevanz von 6 zugewiesen wurde. Auf Grund dessen und , weil sich auch Annotationen mit niedrigen E-values in dieser Relevanzgruppe befinden können, wenn sie beispielsweise nur in

TrEMBL gefunden wurden, stellen die Annotationen aus der niedrigsten Gruppe trotzdem eine wichtige Informationsquelle dar. Sie können für die Modellierer eines metabolischen Netzwerks als Hinweise dienen, wo eine Enzymfunktion im Genom eventuell lokalisiert sein könnte. Auch kann es Biologen für die Laborarbeit Hinweise geben, welche Gene untersucht werden sollten, da sie eventuell für wichtige Enzymfunktionen kodieren.

Für eine automatische Weiterverarbeitung sind die Ergebnisse der niedrigsten Relevanzgruppe allerdings nicht geeignet. Sie müssen in jedem Fall per Hand geprüft werden, und es muss abgewägt werden, ob das Ergebnis valide ist oder nicht.

Die Ergebnisse der anderen Gruppen können, je nach gewünschter Genauigkeit, automatisch übernommen werden. Durchschnittlich etwa ein Drittel aller Ergebnisse liegt in einem Gesamtrelevanzbereich, wo kein bzw. kaum Zweifel an der Annotation besteht.

Vergleicht man die Werte der neun Referenzorganismen, ist auch hier deutlich zu sehen, dass die Ergebnisse von *E. coli* eine deutlich höhere Qualität haben als die der anderen Organismen. Es können sehr viel mehr Ergebnisse in der besten Gruppe erzielt werden. Da *E. coli* sehr gut untersucht ist, finden sich auch in jeder der Annotationsquellen viele Informationen zu diesem Organismus. Somit kommt der Fall, dass mehrere Quellen das gleiche Gen mit der gleichen Funktion annotieren, häufiger vor, und dies führt zu hohen Gesamtrelevanzen.

3.2 Operonanalyse

3.2.1 Länge der Operons

Für die Auswertung der Operons ist ihre Länge entscheidend. In Abbildung 12 ist eine Übersicht über die Längen der Operons gemittelt über die 90 untersuchten Organismen gezeigt.

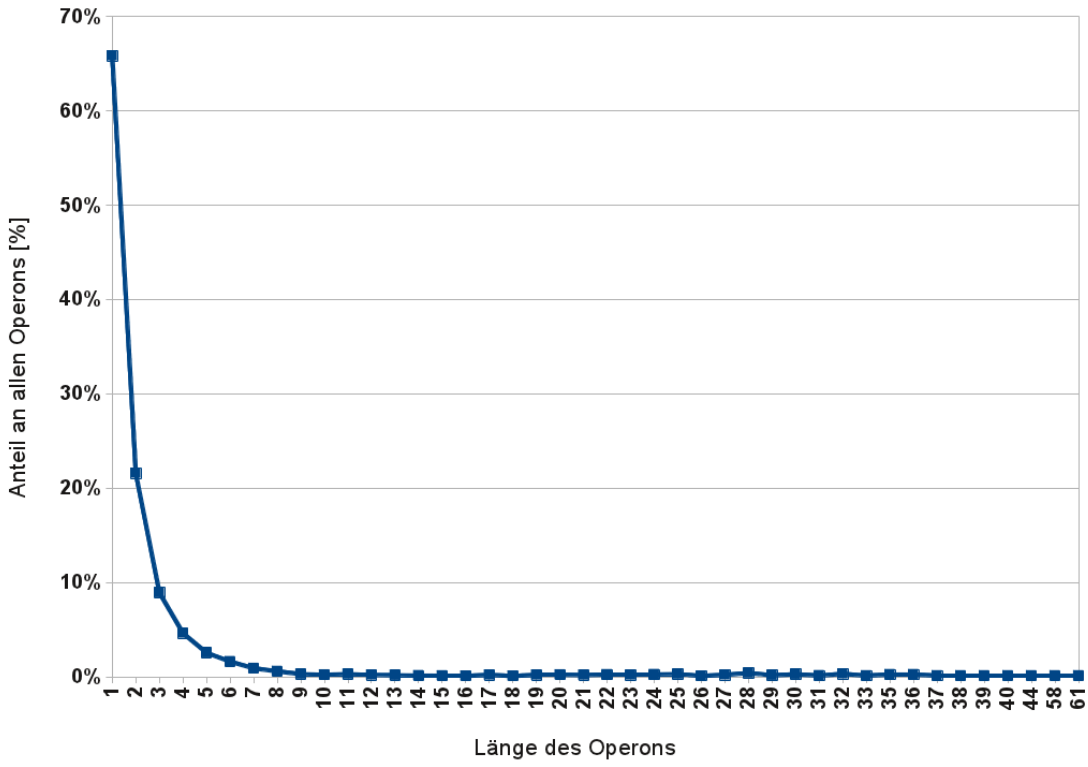


Abbildung 12 - Übersicht über die Längenverteilung der DOOR-Operons der 90 untersuchten Organismen

Es ist deutlich zu sehen, dass der Großteil der Operons in einem Längenbereich bis drei vorkommt. Über 65 % der Operons der Datenbank DOOR haben nur eine Länge von eins, d.h. sie enthalten nur ein Gen und werden in die Analyse nicht einbezogen. Da bei diesen Operons kein Zusammenhang zwischen mehreren Genen hergestellt wird, liefern sie keinen Zugewinn an Information.

3.2.2 Anteil der Gene, die von der Operonanalyse betroffen sind

Nur für einen kleinen Teil der Gene, für die durch die BLAST-Suche bzw. durch eine der anderen Annotationsquellen die Zuweisung einer Enzymfunktion möglich war, wird in der Operonanalyse ein Wert zugewiesen. In Tabelle 11 ist der Anteil dieser Gene an den annotierten Genen dargestellt. Außerdem ist der Anteil der Gene markiert, für die eine Swiss-Prot Annotation zu finden war.

Tabelle 11 - Anzahl der annotierten Gene und der Anteil, der durch die Operonanalyse betroffenen, der in Swiss-Prot annotierten bzw. der Operon betroffenen und Swiss-Prot annotierten an den insgesamt annotierten Genen

Organismus	**annotierte Gene**	**Anteil Operon**	**Anteil Swiss-Prot**	**Anteil Operon/ Swiss-Prot**
C. glutamicum	2204	9 %	14 %	4 %
D. shibae	2699	10 %	6 %	1 %
E. coli	6125	8 %	24 %	4 %
P. aeruginosa	4842	9 %	14 %	3 %
P. putida	4310	9 %	10 %	2 %
S. solfataricus	1837	11 %	14 %	3 %
T. thermophilus	2061	11 %	11 %	3 %
Y. pseudotuberculosis IP32953	3150	8 %	22 %	2 %
Y. pseudotuberculosis YPIII	2798	9 %	12 %	2 %
Durchschnitt der Referenzorganismen		9 %	14 %	3 %

Die Werte zeigen, dass der Anteil der Swiss-Prot Annotationen zwar abhängig davon ist, wie gut ein Organismus untersucht ist, dies die Operonanalyse aber nicht betrifft. Für diese ist der Anteil der betroffenen Gene für alle Referenzorganismen konstant mit minimalen Abweichungen. Diese Tatsache macht die Daten aus der Operonanalyse besonders für nicht so gut untersuchte Organismen interessant, da auch hier für ca. 10 % der annotierten Gene eine Entscheidungshilfe durch die Operonanalyse möglich ist.

Generell muss aber herausgestellt werden, dass der Anteil der schon annotierten Gene, die von der Operonanalyse betroffen sind, mit 10 % nicht sehr hoch ist. Der Anteil der Gene, die von der Operonanalyse betroffen sind und gleichzeitig eine Annotation in Swiss-Prot haben, ist noch geringer. Dieser liegt durchschnittlich bei nur 3 %.

Diese Daten unterstreichen die Schwierigkeit, die Operonanalyse als Annotationsquelle einzuordnen. Die Überschneidung mit dem Wahrheitsstandard ist zu gering, um die Operonanalyse-Ergebnisse durch Vergleich mit diesem einzuordnen.

3.2.3 Güte der Daten aus der Operonanalyse

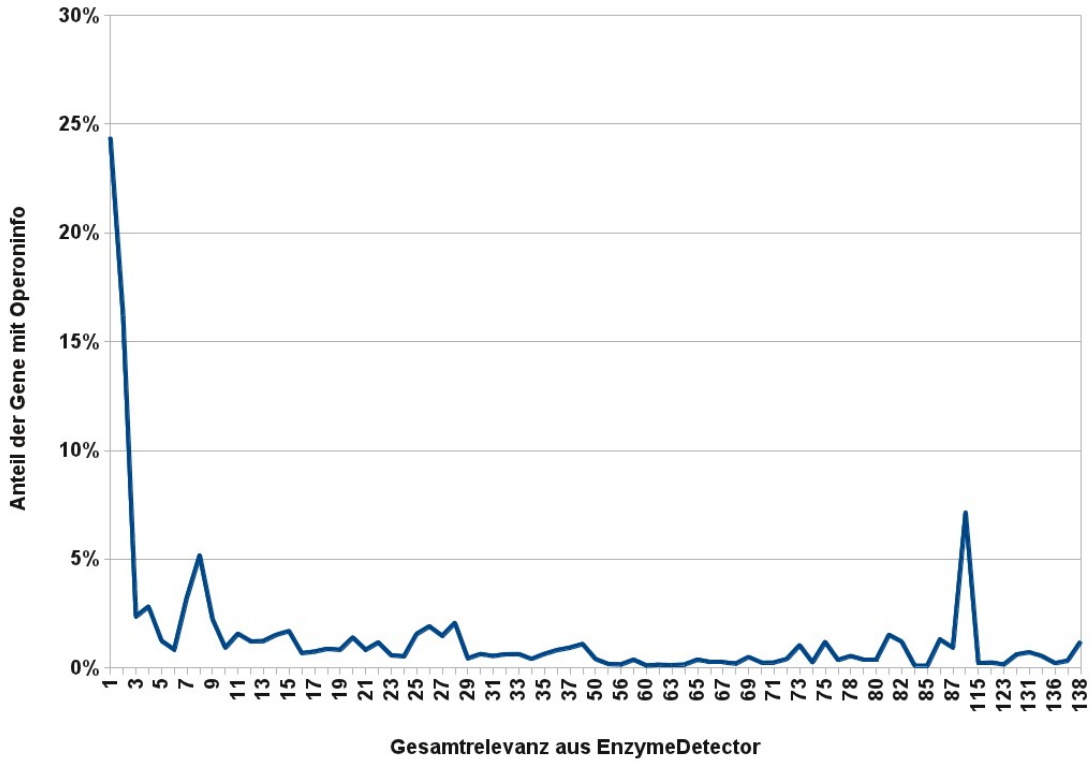


Abbildung 13 - Daten aus der Operonanalyse aufgetragen gegen die Relevanz mit der diese Annotationen im EnzymeDetector gefunden wurden (Durchnittswerte der Referenzorganismen) (Abbildung groß unter Anhang 5)

Für die 3 % der annotierten Gene mit Überschneidung zwischen Swiss-Prot und der Operonanalyse liegt die *Precision* durchschnittlich bei 45 %, und der *Recall* liegt bei 20 %. Dies führt zu einem durchschnittlichen F1-Maß von 29 %. Diese Werte zeigen, dass die Daten nicht geeignet sind, um alleine eine Genomannotation zu stellen.

In Abbildung 13 sind die Daten aus der Operonanalyse gegen die Gesamtrelevanzwerte aus dem EnzymeDetector aufgetragen. Es handelt sich um die Durchschnittswerte der Referenzorganismen. Es ist deutlich zu sehen, dass ein Großteil der Annotationen der Gene, die von der Operonanalyse betroffen sind, im unteren Relevanzbereich liegt. Dies unterstreicht weiter die Aussage, dass die Operonanalyse alleine nicht als vollwertige Annotationsquelle gewertet werden sollte.

Zu Begründen ist dieses Auftreten von vielen Ergebnissen mit niedrigen Relevanzen durch die Tatsache, dass die Operonanalyse ja untersucht, welche der gefundenen Annotationen der Operons in gemeinsamen Stoffwechselwegen vorkommt. Somit bezieht die Operonanalyse alle Ergebnisse, auch die nicht so guten, in ihre Berechnungen mit ein.

3.2.4 Einbeziehung der Operonanalyse als Entscheidungshilfe

Wie deutlich herausgestellt wurde, eignen sich die Daten der Operonanalyse alleine nicht für eine eigenständige Genomannotation. Möglicherweise ließe sich die Qualität der Analyseergebnisse dadurch verbessern, dass nur Daten ab einer bestimmten Güte in die Analyse einbezogen werden. Das Problem der geringen Überschneidung mit den Swiss-Prot Daten bliebe aber bestehen bzw. würde sich noch verstärken und somit würde sich zwar mutmaßlich die Qualität der Ergebnisse der Operonanalyse verbessern, diese aber genau zu bestimmen, würde schwieriger werden.

Daher haben wir uns dazu entschieden, vorerst die Ergebnisse der Operonanalyse nur bei Fällen einzubeziehen, bei denen sonst keine klare Entscheidung möglich war, also für Fälle, bei denen mehrere Kandidaten nach der BLAST-Auswertung zur Verfügung standen.

In Tabelle 12 ist die Anzahl dieser Fälle in den Referenzorganismen aufgezeigt. Zusätzlich ist die Gesamtzahl der Fälle aufgeführt, für die bei der BLAST-Auswertung keine klare Entscheidung getroffen werden konnte.

Es ist deutlich, dass zwar nur für durchschnittlich 13 % der bis hier unentscheidbaren Fälle der Lösungsraum mit Hilfe der Operonanalyse eingeschränkt werden kann, aber

auch damit ist man wieder einen Schritt weiter und bietet dem Nutzer einen Hinweis mehr, welche der Annotationskandidaten am wahrscheinlichsten sind.

Tabelle 12 - Anzahl der Gene der Referenzorganismen, für die keine eindeutige Entscheidung bei der BLAST-Auswertung möglich war und prozentualer Anteil davon, bei dem die Operonanalyse als Entscheidungshilfe fungieren kann

Organismus	**Anzahl Gene mit uneindeutiger BLAST-Auswertung**	**Anteil der Gene bei dem die Operonanalyse Entscheidungshilfe sein kann**
C. glutamicum	491	8 %
D. shibae	733	12 %
E. coli	1173	25 %
P. aeruginosa	1216	13 %
P. putida	1120	12 %
S. solfataricus	319	16 %
T. thermophilus	419	14 %
Y. pseudotuberculosis IP32953	661	10 %
Y. pseudotuberculosis YPIII	645	11 %

Die Daten der Operonanalyse sind noch nicht auf dem Webinterface verfügbar, werden aber in Zukunft auch dort zur Verfügung gestellt.

3.2.5 Fallbeispiele

Wie schon erwähnt, sind wenig Fälle pro Organismus vorhanden, bei denen die BLAST-Auswertung kein klares Ergebnis liefern konnte, die aber von der Operonanalyse betroffen sind.

Ein Beispiel für einen solchen Fall ist das Gen mit der GI 62389059 aus *C. glutamicum*. In der BLAST-Suche wurden zwei mögliche Annotationskandidaten gefunden: das Enzym 5.3.1.1 und das Enzym 4.2.99.18. Beide wurden mit einem E-value von ca. 10^{-60} in TrEMBL gefunden und erhielten so die Relevanz 2. Allerdings wurde nur das Enzym

5.3.1.1 durch die Operonanalyse bestätigt. Somit ist anhand der Operon-Daten hier eine Entscheidung möglich, wo dies vorher nicht möglich war.

Leider gibt es aber auch den Fall, dass auch mit Hilfe der Operonanalyse keine klare Entscheidung möglich ist, sondern der Lösungsraum nur eingeschränkt wird. Beispiel hierfür ist das Gen mit der GI 62389079 aus *C. glutamicum*. Aus dem EnzymeDetector erhält man drei Annotationskandidaten für dieses Gen: Die Enzyme 1.4.7.1, 1.4.1.13 und 1.4.1.14. Von diesen drei Enzymen werden aber nur zwei durch die Operonanalyse bekräftigt, und zwar die Enzyme 1.4.7.1 und 1.4.1.14.

Somit wird durch die Operonanalyse die Glutamatsynthase bestärkt, die mit Hilfe von NADH agiert (1.4.1.14) und nicht die, die mit NADPH (1.4.1.13) agiert.

Es bleibt festzustellen, dass mit der aktuellen Vorgehensweise die Operonanalyse nur eine kleine Hilfe für den Teil der Gene ist, bei denen noch keine klare Entscheidung für einen Annotationskandidaten getroffen werden konnte.

Um die Qualität der Ergebnisse zu verbessern und somit die Ergebnisse möglichst als eigenständige Genomannotation werten zu können, müssen weitere Untersuchungen durchgeführt werden.

3.3 EnzymeDetector Webinterface

Das Webinterface des EnzymeDetector ist in eine Hauptseite und vier Darstellungsformen der Ergebnisse unterteilt, die über einzelne Tabs angesteuert werden können

3.3.1 Hauptseite

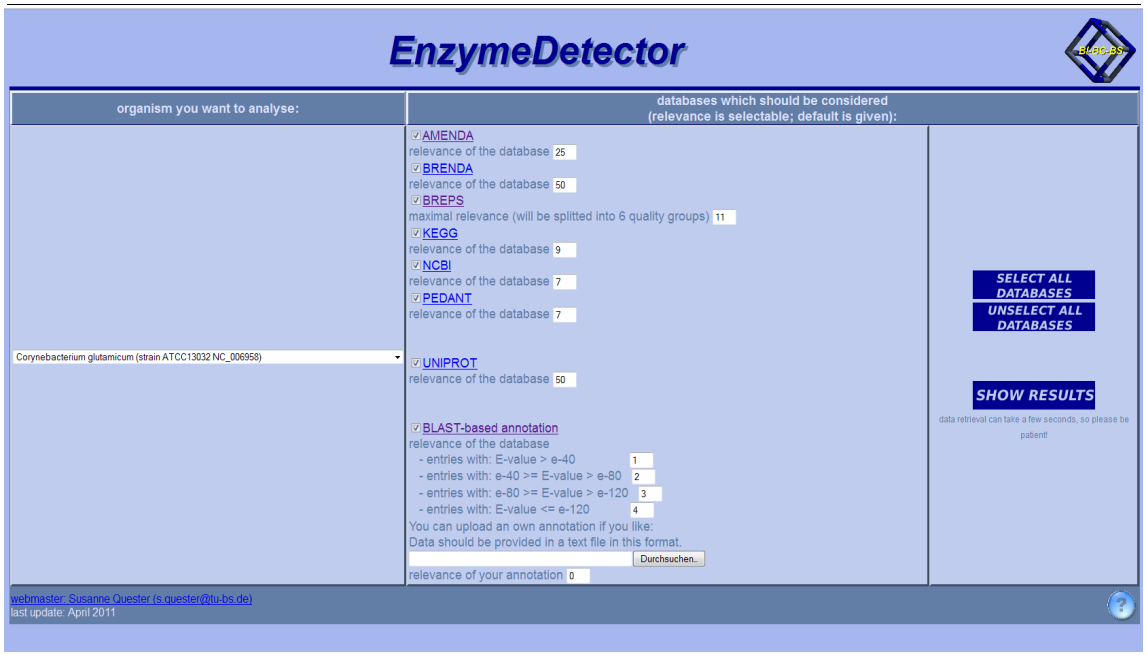


Abbildung 14 - Hauptseite des EnzymeDetector-Webinterface

Auf der Eingangsseite des Webinterfaces (siehe Abbildung 14) hat der Nutzer die Möglichkeit, mittels eines Aufklapp-Menüs einen Organismus auszuwählen. Die wählbaren Organismen werden hierfür aus der „Species“-Tabelle abgefragt.

Nachdem der Organismus ausgewählt wurde, erscheinen die Annotationsquellen, die für diesen Organismus verfügbar sind. Diese Information wird aus der „Input_Databases“-Tabelle extrahiert und auf der Webseite bereitgestellt. In dieser Tabelle finden sich auch die Standardwerte der Relevanzen für die einzelnen Annotationsquellen. Es können entweder alle Quellen oder gezielt einzelne Quellen selektiert werden.

Der Nutzer hat die Möglichkeit, diesen Quellen eigene Einzelrelevanzen zuzuweisen. Es werden die unter Punkt 2.12 auf Seite 34 beschriebenen Standardwerte vorgegeben. Diese können jederzeit geändert werden.

Der Nutzer hat an dieser Stelle zusätzlich die Möglichkeit, eine eigene Genomannotation hochzuladen, um sie mit den ausgewählten Quellen zu vergleichen. Diese Nutzer-Genomannotation wird im Folgenden wie jede andere Annotationsquelle behandelt, also muss auch ihr eine Einzelrelevanz durch den Nutzer zugewiesen werden.

Die ausgewählten Parameter werden durch Klicken auf den „Show results“-Knopf übernommen, und die Daten werden vom Server abgefragt. Dieser Schritt kann auf Grund des Datenvolumens einige Sekunden in Anspruch nehmen.

3.3.2 Standard Grenzwerte

Zusätzlich zu den Einzelrelevanzen, die der Nutzer, falls gewünscht, selbst zuweisen kann, werden die Daten auch nach Standard-Grenzwerten gefiltert. So werden nicht-relevante Daten ausgeblendet und so die Ergebnisansicht übersichtlicher. Die Grenzwerte können jederzeit angepasst werden.

Zum einen gibt es einen Grenzwert für die Einbeziehung der Daten in die BLAST-basierte Genomannotation. Standardmäßig werden nur BLAST-Hits als mögliche Annotationskandidaten berücksichtigt, wenn der zugehörige beste E-value kleiner als 10^{-25} ist. Dieser Grenzwert kann vom Nutzer nach eigenem Ermessen geändert werden.

Des Weiteren werden die Daten nach einem Gesamtrelevanz-Grenzwert gefiltert. Ein Grenzwertschema mit drei verschiedenen Grenzwerten wurde ermittelt. Hierfür wurde eine Untersuchung der *Precision*, des *Recalls* und des F1-Maßes der Annotationsergebnisse aller 81 untersuchten Organismen im Vergleich zu den Annotationsergebnissen aus Swiss-Prot durchgeführt. Die Ergebnisse der neun Organismen des Trainingsdatensatzes wurden von dieser Analyse ausgeschlossen, da mit ihnen die Standardwerte festgelegt wurden und für die vorliegende Analyse eine Trennung der Trainings- und Testdaten nötig war. Die Ergebnisse der Untersuchung sind in Abbildung 15 zu sehen.

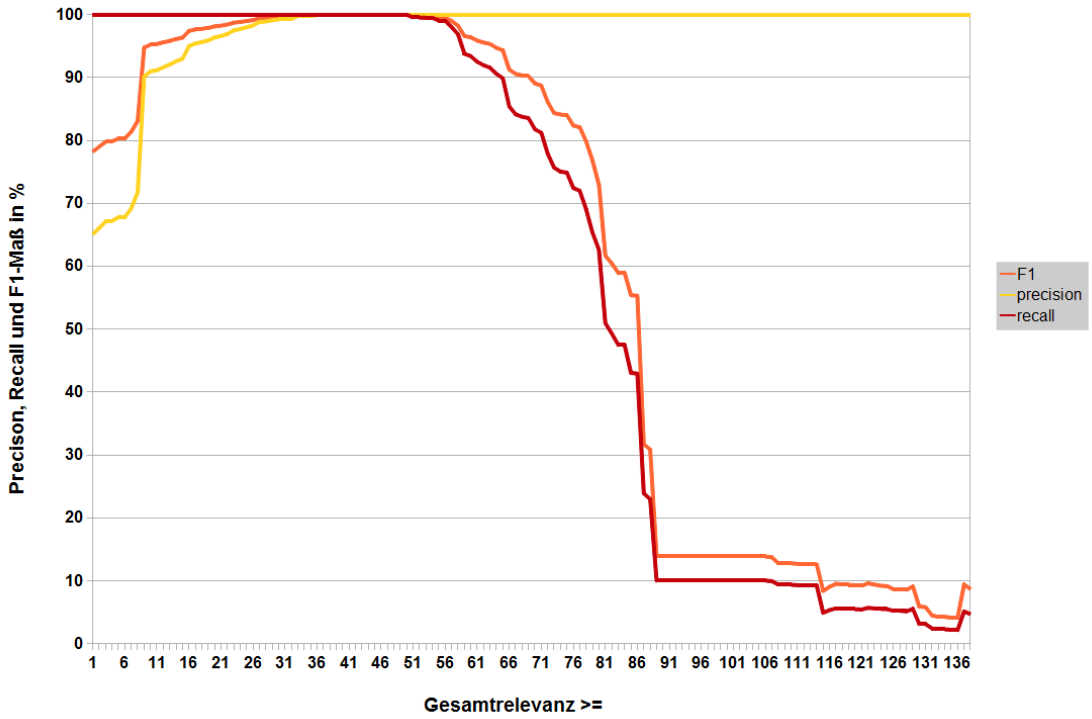


Abbildung 15 - F1-Maß, Precision und Recall der Ergebnisse des EnzymeDetectors im Vergleich zur Swiss-Prot Genomannotation aufgetragen gegen die Gesamtrelevanz

Für eine großzügige Filterung der Daten wird ein Grenzwert von 7 für die Gesamtrelevanz vorgeschlagen. Dieser Wert ist auch als Standardwert für das Webinterface hinterlegt. Mit diesem Grenzwert werden Daten mit einer optimalen Trefferquote erhalten, allerdings zu Lasten der Genauigkeit der Ergebnisse. Der Vorteil dieses relativ niedrigen Grenzwertes besteht darin, dass so Annotationen von Genen, die bisher noch nicht in einer der Datenbanken annotiert waren, aber in die BLAST-basierte Genomannotation aufgenommen wurden, nicht verloren gehen. Wie im Ergebnisteil dieser Arbeit dargestellt, liefert die BLAST-basierte Genomannotation einen enormen Zugewinn an Daten. Würde man nur mit Annotationen arbeiten, die in mehreren Quellen gleichzeitig vorkommen, wäre die gefundene Information zwar verlässlicher, als wenn sie nur in einer Quelle vorkommt, allerdings verschließt man sich so dem Zugewinn von Daten, die auf neuannotierten Sequenzen beruhen. Somit ist der niedrige Grenzwert des Schemas dann anzuraten, wenn es dem Nutzer darum geht, alle Enzyme des Organismus zu erhalten, auch wenn das Ergebnis eventuell zusätzliche Enzyme enthält, die nicht der Wahrheit entsprechen.

Für eine mittelstrikte Filterung wird ein Grenzwert von 26 vorgeschlagen. Dies ist der niedrigste Gesamtrelevanzwert, für den das F1-Maß der Ergebnisse über 99 % liegt. Der Nutzer erhält mit diesem Grenzwert Daten mit einer perfekten Trefferquote und sehr guter Genauigkeit.

Wenn die Daten nach möglichst hoher Genauigkeit gefiltert werden sollen, ist ein Grenzwert von 39 anzuraten. Dies ist der niedrigste Gesamtrelevanzwert für den das *F1-Maß* maximal ist. Das heißt, der Nutzer erhält eine hohe Genauigkeit ohne den Verlust von Informationen.

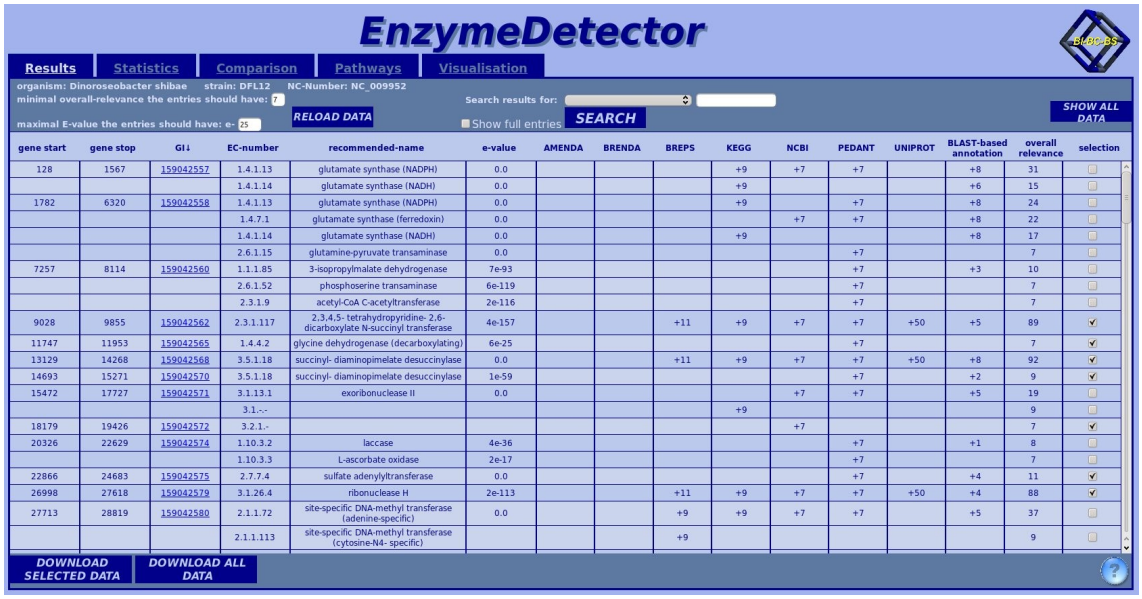
Es gilt zu beachten, dass für die Bestimmung dieser Grenzwerte nur Gene einbezogen werden konnten, die auch in Swiss-Prot annotiert sind. Der Anteil dieser Gene liegt für alle Organismen durchschnittlich bei 14 %.

Die Daten sprechen zwar dafür, immer einen Grenzwert von 39 anzuwenden, da das F1-Maß hier ja maximal ist. Allerdings muss beachtet werden, dass die Gene, die in Swiss-Prot annotiert sind, recht bekannte Gene sind, die meist auch in allen anderen Datenbanken annotiert sind. In den meisten Fällen werden so hohe Gesamtrelevanzen für relativ neusequenzierte, neuannotierte und unbekannte Gene nicht erreicht werden. Durch einen so hohen Grenzwert würde man also hier die komplette Information verlieren. Allerdings gibt es Anwendungen, bei denen es mehr darum geht, sehr verlässliche Informationen zu erhalten, als alle möglichen Enzyme. In diesen Fällen sind hohe Grenzwerte angebracht.

3.3.3 Ergebnis-Seite

Nach Erfragen der Daten vom Server werden diese standardmäßig zunächst in der tabellarischen Darstellung angezeigt. Im oberen Bereich der Seite finden sich Tabs, mit denen der Nutzer zwischen den verschiedenen Ansichten der Daten wechseln kann. Bei Tabs handelt es sich um Elemente, die der Aufteilung der Seite dienen. Sie sind den Registrierkarten aus Aktenschränken nachempfunden.

Unter den Tabs befindet sich die eigentliche Datenseite. Diese besteht aus einer Kopfzeile und den Daten in der aktuell ausgewählten Ansicht. Die Ergebnisseite ist in Abbildung 16zu sehen.



gene start	gene stop	GI↓	EC-number	recommended-name	e-value	AMENDA	BRENDA	BREPS	KEGG	NCBI	PEDANT	UNIPROT	BLAST-based annotation	overall relevance	selection
128	1567	159042557	1.4.1.13	glutamate synthase (NADPH)	0.0				+9	+7	+7		+8	31	
			1.4.1.14	glutamate synthase (NADH)	0.0				+9				+6	15	
1782	6320	159042558	1.4.1.13	glutamate synthase (NADPH)	0.0				+9		+7		+8	24	
			1.4.7.1	glutamate synthase (ferredoxin)	0.0					+7	+7		+8	22	
			1.4.1.14	glutamate synthase (NADH)	0.0				+9				+8	17	
			2.6.1.15	glutamine-pyruvate transaminase	0.0						+7			7	
7257	8114	159042560	1.1.1.85	3-isopropylmalate dehydrogenase	7e-93						+7		+3	10	
			2.6.1.52	phosphoserine transaminase	6e-119						+7			7	
			2.3.1.9	acetyl-CoA C-acetyltransferase	2e-116						+7			7	
9028	9855	159042562	2.3.1.117	2,3,4,5- tetrahydropyridine- 2,6-dicarboxylate N-succinyl transferase	4e-157			+11	+9	+7	+7	+50	+5	89	☑
11747	11953	159042565	1.4.4.2	glycine dehydrogenase (decarboxylating)	6e-25						+7			7	☑
13129	14268	159042568	3.5.1.18	succinyl- diaminopimelate desuccinylase	0.0			+11	+9	+7	+7	+50	+8	92	☑
14693	15271	159042570	3.5.1.18	succinyl- diaminopimelate desuccinylase	1e-59						+7		+2	9	☑
15472	17727	159042571	3.1.13.1	exoribonuclease II	0.0					+7	+7		+5	19	
			3.1.-.-						+9					9	
18179	19426	159042572	3.2.1.-							+7				7	☑
20326	22629	159042574	1.10.3.2	laccase	4e-36						+7		+1	8	
			1.10.3.3	L-ascorbate oxidase	2e-17						+7			7	
22866	24683	159042575	2.7.7.4	sulfate adenylyltransferase	0.0						+7		+4	11	☑
26998	27618	159042579	3.1.26.4	ribonuclease H	2e-113			+11	+9	+7	+7	+50	+4	88	☑
27713	28819	159042580	2.1.1.72	site-specific DNA-methyl transferase (adenine-specific)	0.0			+9	+9	+7	+7		+5	37	
			2.1.1.113	site-specific DNA-methyl transferase (cytosine-N4- specific)				+9						9	

Abbildung 16 - Ergebnisseite des EnzymeDetector-Webinterfaces

In der Kopfzeile erhält der Nutzer Informationen zum gewählten Organismus und den Grenzwerten, die zur Filterung der Daten verwendet wurden. In diesem Bereich können diese Grenzwerte auch jederzeit geändert werden.

In der Mitte der Kopfzeile befindet sich ein Suchmenü, das es dem Nutzer ermöglicht, die gesamten Daten nach bestimmten Kriterien zu durchsuchen. Die Daten können nach einem bestimmten Genstart oder -stop, einer GI, dem Vorkommen einer EC-Nummer oder einem Enzymnamen durchsucht werden. Außerdem ist es möglich, sich nur die Annotationen anzeigen zu lassen, die in bestimmten Quellen vorkommen bzw. einer bestimmten Kombination von Quellen. Für alle Suchkriterien ist eine Gruppen- und eine Ähnlichkeitssuche implementiert. Es können also mehrere Suchbegriffe des gleichen Kriteriums auf einmal an die Suche übergeben werden. Darüber hinaus ist es möglich, mit dem Symbol „*“ als Platzhalter zu arbeiten, falls ganze Gruppen von Ergebnissen gefunden werden sollen.

Des Weiteren ist es möglich, sich für jeden Suchtreffer auch alle alternativen Annotationen für denselben Locus anzeigen zu lassen. Diese Option wird durch ein

Häkchen im Feld „Show full entries“ aktiviert. Diese Option erweist sich als sinnvoll, wenn zum Beispiel nach einer bestimmten EC-Nummer im Genom gesucht wird, man aber auch die alternativen EC-Kandidaten für den Locus kennen möchte, um die Annotation wirklich zu beurteilen.

Ganz rechts in der Kopfzeile befindet sich der „Show all data“-Knopf, mit dem alle Suchen zurückgesetzt werden können und wieder eine reine Filterung nach den Grenzwerten eingestellt wird.

Im eigentlichen Datenteil der Seite findet sich eine tabellarische Auflistung der gefilterten EnzymeDetector-Ergebnisse. Hier findet sich ein Eintrag für jede Gen-Enzym-Kombination, die durch das Programm gefunden wurde, und die nicht genspezifischen Vorhersagen. Für den Fall, dass mehrere Annotationskandidaten die Filterung erfüllen, werden alle untereinander aufgelistet. Sortiert sind die Daten standardmäßig nach der Genposition, aber sie können auch nach der GI, dem Standardnamen des Enzyms oder der EC-Nummer sortiert werden. Dies geschieht einfach durch Klicken auf den entsprechenden Spaltentitel.

Zu jeder Annotation finden sich Informationen zum Gen, wie Genstart, Genstop und GI, und Informationen zur Annotation selbst, wie EC-Nummer, Standardname des Enzyms, E-value, mit dem die EC-Nummer in der BLAST-Suche gefunden wurde, und Informationen dazu, welche Quellen die Annotation bestätigen. Ist kein E-value eingetragen, so heißt dies, dass die Annotation zwar in einer der Quellen gefunden wurde, allerdings nicht im Rahmen der BLAST-Suche.

Wird die Annotation durch eine der Quellen bestätigt, so findet sich in der zugehörigen Spalte die Einzelrelevanz dieser Quelle. Bestätigt die Quelle die Annotation nicht, so bleibt das Feld leer. In der vorletzten Spalte der Tabelle findet sich die Gesamtrelevanz, die es dem Nutzer erleichtern soll, sich für eine Annotation zu entscheiden, falls es mehrere Kandidaten gibt.

Mit der letzten Spalte der Tabelle hat der Nutzer die Möglichkeit, bestimmte Annotationen auszuwählen. Gibt es nur einen Annotationskandidaten für ein Gen, ist

diese Annotation vorselektiert. Gibt es mehrere, so kann der Nutzer die Entscheidung treffen, welche Annotation(en) er in die Auswahl übernehmen möchte.

Unten links auf der Seite hat der Nutzer die Möglichkeit, entweder den gesamten Datensatz oder die von ihm getroffene bzw. die vorselektierte Auswahl an Daten als CSV-Datei herunterzuladen. Dies ermöglicht eine einfache Weiterbearbeitung der Daten durch den Nutzer. CSV ist, wie andere ASCII-Formate auch, sehr praktisch für den Datentausch zwischen Programmen und wird von allen Tabellenkalkulationsprogrammen unterstützt.

3.3.4 Statistik-Seite

Durch Klick auf das Tab „Statistics“ gelangt der Nutzer auf die Seite mit der statistischen Auswertung der Daten (siehe Abbildung 17). Auch diese Seite besteht aus der Kopfzeile und einem Datenbereich. Neben den Informationen zum Organismus findet sich in der Kopfzeile die Information, wie viele Gene der Organismus hat. Diese Information wird per SQL-Abfrage aus dem Zusammenschluss aller Datenbank-Informationen entnommen.

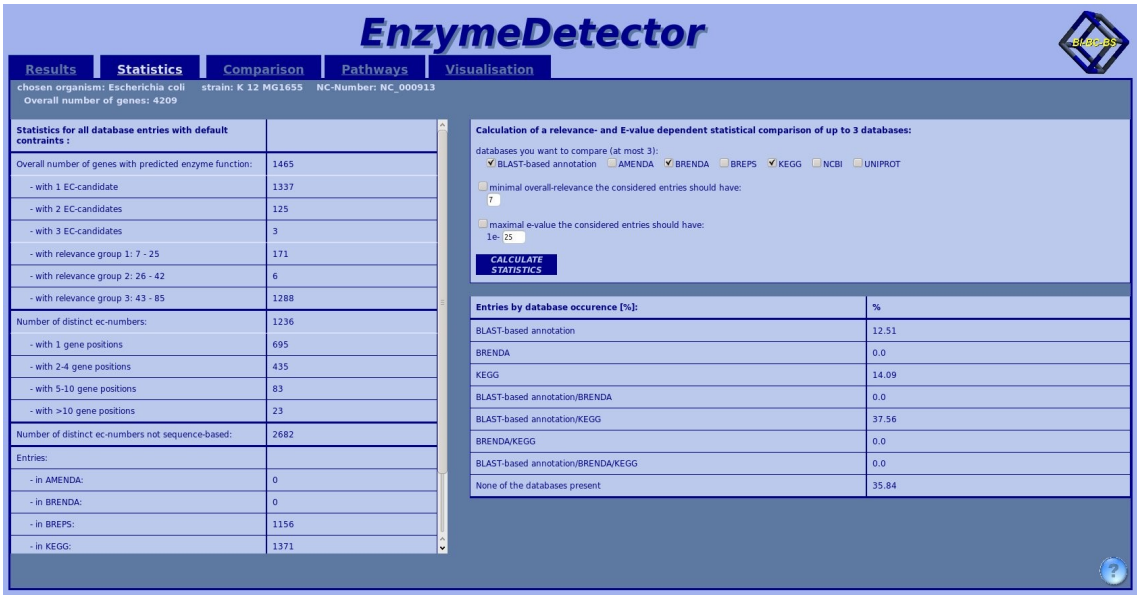


Abbildung 17 - Statistikseite des EnzymeDetector-Webinterfaces

Die Statistikseite besteht aus zwei Teilen, dem statischen und dem dynamischen Teil. Dies ermöglicht dem Nutzer, die Auswirkung von verschieden gewählten Grenzwerten

auf einfache Art und Weise zu untersuchen. Im linken Teil der Seite befinden sich die Statistikwerte für die Standardgrenzwerte (maximaler E-value 10^{-25}; minimale Gesamtrelevanz 7). Im rechten Teil die gleichen Werte, berechnet mit ausgewählten Grenzwerten. Somit können die statistischen Werte unterschiedlicher Grenzwerte einfach miteinander verglichen werden. Der Nutzer kann so auch ermitteln, wie viele Daten und welche Art von Daten bei welchem Grenzwert verloren gehen.

Der dynamische Teil der Seite ermöglicht es weiterhin, bis zu drei Annotationsquellen miteinander zu vergleichen. Es wird ermittelt, inwiefern die Quellen in ihren Annotationen übereinstimmen und wie viel Prozent der Annotationen nur von einer Quelle geliefert werden.

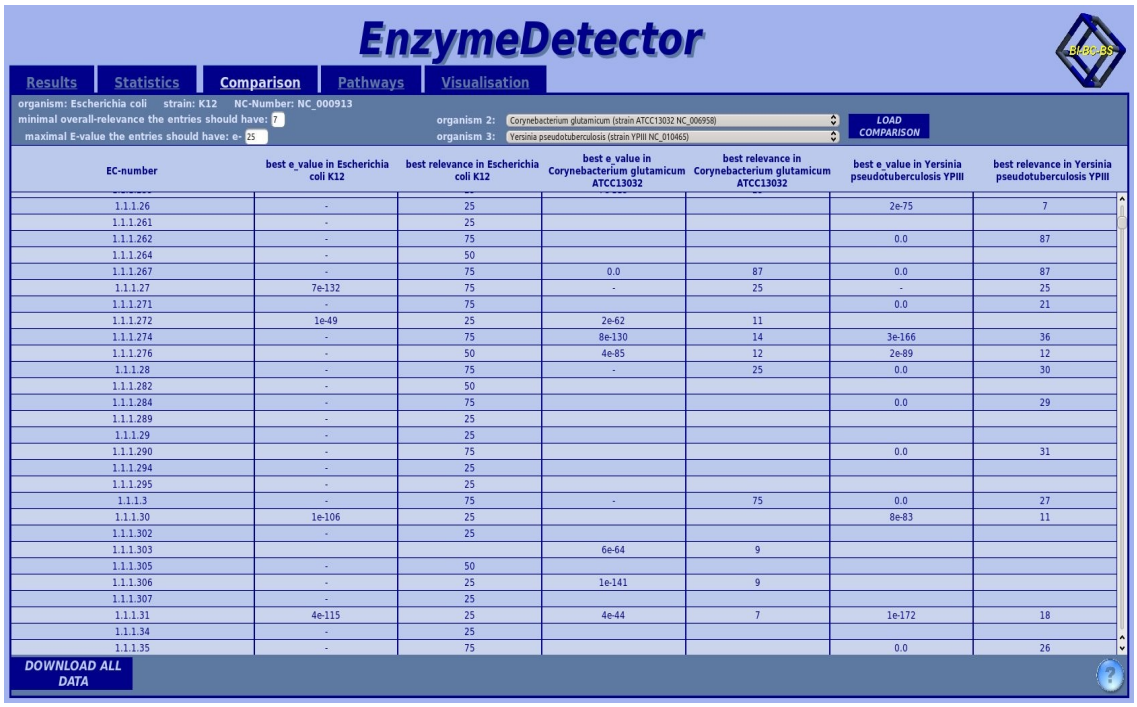
Aufgeführt sind u.a. die Anzahl der Gene, die für Enzyme kodieren, die Anzahl der Annotationen in den verschiedenen Gesamtrelevanzbereichen und die Anzahl der verschiedenen EC-Nummern, die für den Organismus gefunden wurden.

Auch die Anzahl der Annotationen, die nicht sequenzbasiert gefunden wurden, wird angegeben. Außerdem erfährt der Nutzer, wie viele der Einträge aus der Ergebnistabelle von welchen Annotationsquellen vorhergesagt werden.

3.3.5 Vergleichs-Seite

Die Vergleichsseite bietet dem Nutzer die Möglichkeit, die gesamte Liste des Enzymbestands des untersuchten Organismus nach EC-Nummern sortiert zu sehen. Es können außerdem bis zu zwei weitere Organismen ausgewählt werden, deren Enzymbestand mit dem des untersuchten Organismus verglichen wird. Der Nutzer gelangt auf diese Ansicht der Daten, indem er auf das „Comparison“-Tab klickt.

Für jede Enzymfunktion ist aufgelistet, mit welchem E-value und mit welcher Gesamtrelevanz sie im Organismus gefunden wurde. Es zählen hierfür jeweils die besten Werte der beiden Kategorien.



EC-number	best e_value in Escherichia coli K12	best relevance in Escherichia coli K12	best e_value in Corynebacterium glutamicum ATCC13032	best relevance in Corynebacterium glutamicum ATCC13032	best e_value in Yersinia pseudotuberculosis YPIII	best relevance in Yersinia pseudotuberculosis YPIII
1.1.1.26	-	25			2e-75	7
1.1.1.261	-	25				
1.1.1.262	-	75			0.0	87
1.1.1.264	-	50				
1.1.1.267	-	75	0.0	87	0.0	87
1.1.1.27	7e-132	75	-	25	-	25
1.1.1.271	-	75			0.0	21
1.1.1.272	1e-49	25	2e-62	11		
1.1.1.274	-	75	8e-130	14	3e-166	36
1.1.1.276	-	50	4e-85	12	2e-89	12
1.1.1.28	-	75	-	25	0.0	30
1.1.1.282	-	50				
1.1.1.284	-	75			0.0	29
1.1.1.289	-	25				
1.1.1.29	-	25				
1.1.1.290	-	75			0.0	31
1.1.1.294	-	25				
1.1.1.295	-	25				
1.1.1.3	-	75	-	75	0.0	27
1.1.1.30	1e-106	25			8e-83	11
1.1.1.302	-	25				
1.1.1.303			6e-64	9		
1.1.1.305	-	50				
1.1.1.306	-	25	1e-141	9		
1.1.1.307	-	25				
1.1.1.31	4e-115	25	4e-44	7	1e-172	18
1.1.1.34	-	25				
1.1.1.35	-	75			0.0	26

Abbildung 18 - Vergleichsseite des EnzymeDetector-Webinterfaces

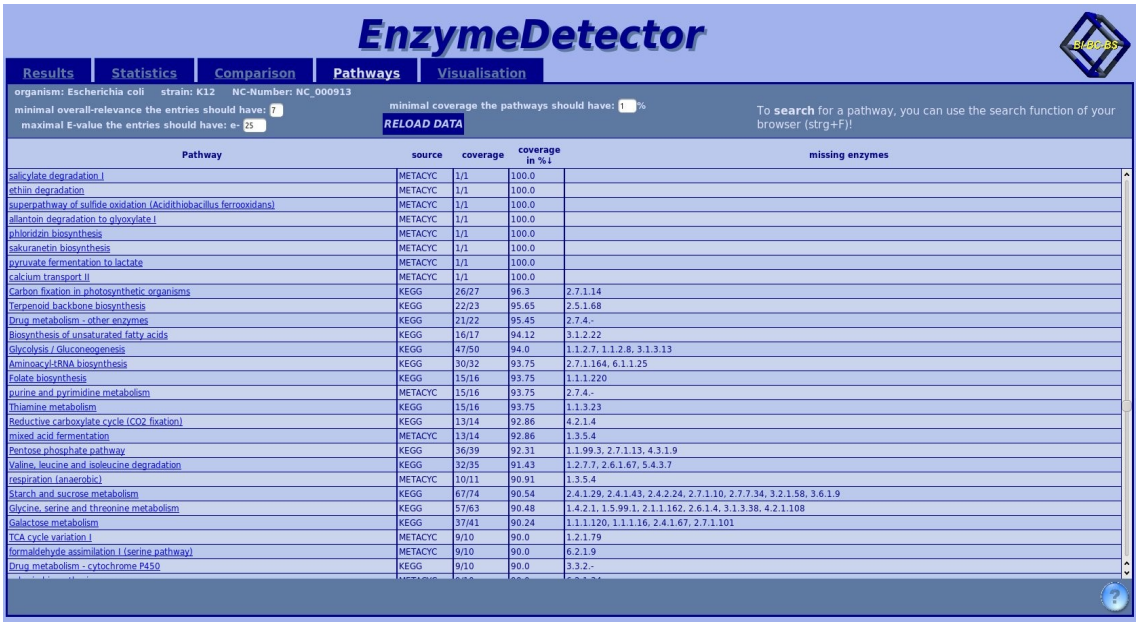
Wird kein E-value anzeigt, ist es entweder der Fall, dass die Enzymfunktion nicht-sequenzbasiert in BRENDA oder AMENDA gefunden wurde und es keinen besseren Hit für diese Funktion gab, oder aber die Enzymfunktion wurde zwar in einer der Annotationsquellen gefunden, aber nicht durch die BLAST-Suche bestätigt. Ein Beispiel der Seite ist in Abbildung 18 zu sehen.

3.3.6 Stoffwechselweg-Seite

Zur Ansicht der Stoffwechselwege gelangt der Nutzer, indem er auf das „Pathways“-Tab in der Tab-Zeile klickt. Man findet hier eine Auflistung aller Stoffwechselwege, die bei KEGG und MetaCyc hinterlegt sind. Zusätzlich erhält man die Information, zu wie viel Prozent dieser Stoffwechselweg im aktuell untersuchten Organismus mit den aktuell gewählten Grenzwerten abgedeckt ist, und ggf. welche Enzyme dem Organismus fehlen, um den Weg vollständig abzudecken.

Die Grenzwerte können jederzeit geändert werden. Außerdem hat der Nutzer die Möglichkeit, sich nur die Stoffwechselwege anzuschauen, die zu einem Mindestprozentsatz abgedeckt sind. Wenn der Nutzer auf den Namen eines

Stoffwechselweges klickt, wird er zur entsprechenden MetaCyc- bzw. KEGG-Seite weitergeleitet, auf der der Weg zu sehen ist.



Pathway	source	coverage	coverage in %↓	missing enzymes
salicylate degradation I	METACYC	1/1	100.0	
ethiin degradation	METACYC	1/1	100.0	
superpathway of sulfide oxidation (Acidithiobacillus ferrooxidans)	METACYC	1/1	100.0	
allantoin degradation to glyoxylate I	METACYC	1/1	100.0	
phloridzin biosynthesis	METACYC	1/1	100.0	
sakuranetin biosynthesis	METACYC	1/1	100.0	
pyruvate fermentation to lactate	METACYC	1/1	100.0	
calcium transport II	METACYC	1/1	100.0	
Carbon fixation in photosynthetic organisms	KEGG	26/27	96.3	2.7.1.14
Terpenoid backbone biosynthesis	KEGG	22/23	95.65	2.5.1.68
Drug metabolism - other enzymes	KEGG	21/22	95.45	2.7.4.-
Biosynthesis of unsaturated fatty acids	KEGG	16/17	94.12	3.1.2.22
Glycolysis / Gluconeogenesis	KEGG	47/50	94.0	1.1.2.7, 1.1.2.8, 3.1.3.13
Aminoacyl-tRNA biosynthesis	KEGG	30/32	93.75	2.7.1.164, 6.1.1.25
Folate biosynthesis	KEGG	15/16	93.75	1.1.1.220
purine and pyrimidine metabolism	METACYC	15/16	93.75	2.7.4.-
Thiamine metabolism	KEGG	15/16	93.75	1.1.3.23
Reductive carboxylate cycle (CO2 fixation)	KEGG	13/14	92.86	4.2.1.4
mixed acid fermentation	METACYC	13/14	92.86	1.3.5.4
Pentose phosphate pathway	KEGG	36/39	92.31	1.1.99.3, 2.7.1.13, 4.3.1.9
Valine, leucine and isoleucine degradation	KEGG	32/35	91.43	1.2.7.7, 2.6.1.67, 5.4.3.7
respiration (anaerobic)	METACYC	10/11	90.91	1.3.5.4
Starch and sucrose metabolism	KEGG	67/74	90.54	2.4.1.29, 2.4.1.43, 2.4.2.24, 2.7.1.10, 2.7.7.34, 3.2.1.58, 3.6.1.9
Glycine, serine and threonine metabolism	KEGG	57/63	90.48	1.4.2.1, 1.5.99.1, 2.1.1.162, 2.6.1.4, 3.1.3.38, 4.2.1.108
Galactose metabolism	KEGG	37/41	90.24	1.1.1.120, 1.1.1.16, 2.4.1.67, 2.7.1.101
TCA cycle variation I	METACYC	9/10	90.0	1.2.1.79
formaldehyde assimilation I (serine pathway)	METACYC	9/10	90.0	6.2.1.9
Drug metabolism - cytochrome P450	KEGG	9/10	90.0	3.3.2.-

Abbildung 19 - Stoffwechselwegseite des EnzymeDetector-Webinterfaces

Ein Beispiel für die Seite ist in Abbildung 19 zu sehen. Standardmäßig ist die Tabelle alphabetisch nach den Stoffwechselwegen sortiert, kann aber auch nach Quelle, d.h. KEGG/MetaCyc, oder Prozentsatz der Abdeckung sortiert werden.

3.3.7 Hilfe-Funktion

Auf jeder der Unterseiten des Webinterfaces ist es möglich, die Hilfe-Funktion dazuzuschalten. Dies geschieht durch Klicken auf das Fragezeichen-Symbol in der rechten unteren Ecke der jeweiligen Seite. Auf gleichem Wege kann diese Funktion auch wieder deaktiviert werden.

Ist die Hilfe-Funktion aktiviert, erhält der Nutzer zusätzliche Informationen zur Verwendung der Seite und zum Objekt, über dem sich der Mauszeiger gerade befindet. Dies geschieht durch die Einblendung von Informationsboxen, wenn der Nutzer mit der Maus über das Objekt fährt.

3.4 Stoffwechselkarte

Wie bereits unter Punkt 2.22 erwähnt, wurde im Rahmen der Arbeit an der Planung und Umsetzung einer generischen Stoffwechselkarte mitgearbeitet. Neben der Planung und Betreuung der Erstellung wurden verschiedene Grafik-Funktionen für die Karte implementiert. Dies geschah zum Teil im Rahmen von Etagenpraktika, die innerhalb der Arbeit betreut wurden. Die unterschiedlichen Grafik-Funktionen wurden mit Hilfe der selbst-vergebenen Attribute der Knoten und Kanten realisiert.

Die Abbildungen, die die verschiedenen Möglichkeiten der Karte darstellen, wurden mit Hilfe von BRIME und Cytoscape erstellt.

3.4.1 Detaillevel

Um eine schnelle Übersicht über die Stoffwechselkarte zu gewährleisten, wurden vier verschiedene Detaillevel realisiert. Der Nutzer kann selbst entscheiden, wie viele der vorhandenen Informationen in der Karte wirklich angezeigt werden sollen.

Im Folgenden sind die verschiedenen Detaillevel beispielhaft anhand des Cyanat-Abbaus dargestellt.

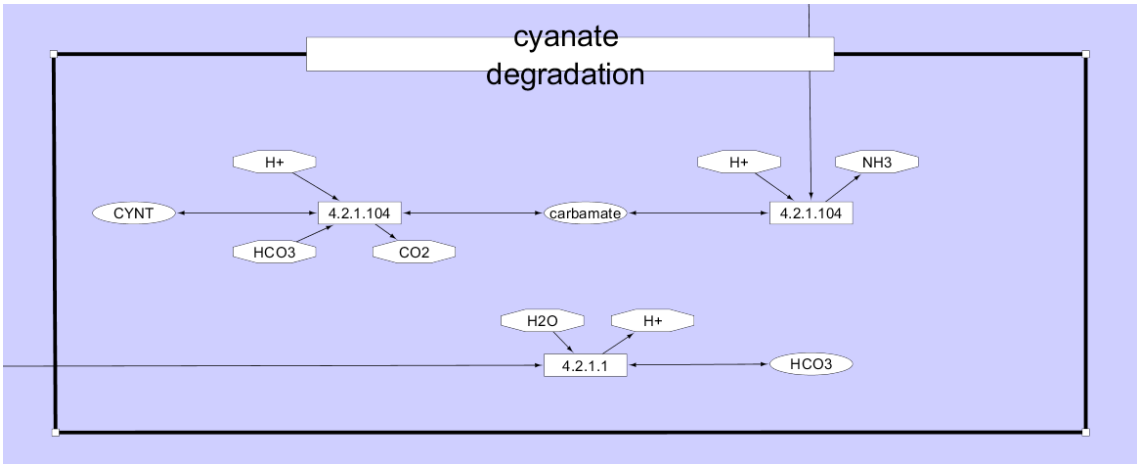


Abbildung 20 - Cyanat-Abbau aus der Stoffwechselkarte dargestellt im höchsten Detaillevel [Abbildung Cytoscape]

Auf dem höchsten Detaillevel der Karte sind alle Knoten und Kanten eingeblendet. Dies ist die Ansicht, in der alle vorhandenen Informationen zu sehen sind und in der die

Karte entwickelt wurde. In Abbildung 20 ist der Cyanat-Abbau auf diesem Detaillevel zu sehen.

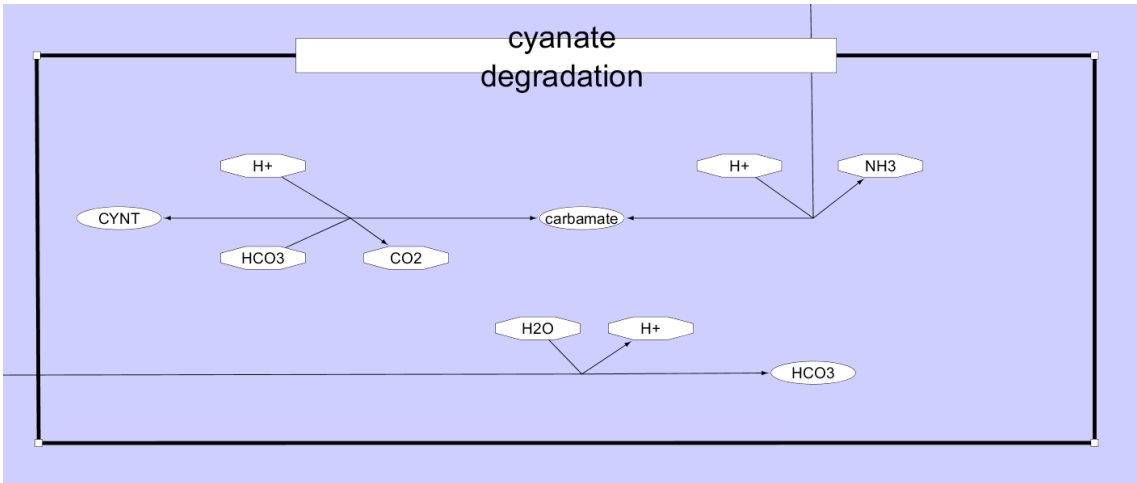


Abbildung 21 - Cyanat-Abbau aus der Stoffwechselkarte im Detaillevel 2: die Enzymknoten sind ausgeblendet [Abbildung Cytoscape]

Im Detaillevel 2 sind die Enzymknoten ausgeblendet, und nur noch die Metabolite sind zu sehen. Auch werden die Kanten so geändert, dass die Pfeilspitzen nur noch auf Metabolite zeigen und nicht auf nicht mehr sichtbare Enzymknoten. Der Cyanat-Abbau in diesem Detaillevel ist in Abbildung 21 zu sehen.

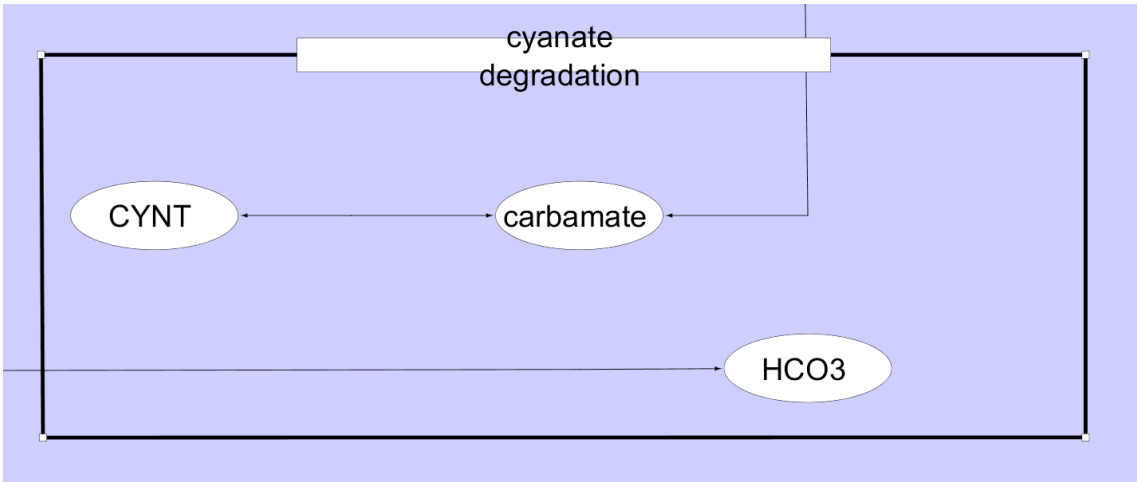


Abbildung 22 - Cyanat-Abbau in Detaillevel 3: Enzymknoten und Nebenmetabolite sind ausgeblendet, und die Größe der Knoten ist angepasst [Abbildung Cytoscape]

Für Detaillevel 3 werden neben den Enzymknoten auch die Nebenmetabolite mit ihren Kanten ausgeblendet. Dies sorgt für eine größere Übersichtlichkeit. Die verbleibenden

Knoten werden in ihrer Größe und der Größe ihrer Beschriftung angepasst. Zu sehen ist dieses Detaillevel in Abbildung 22.

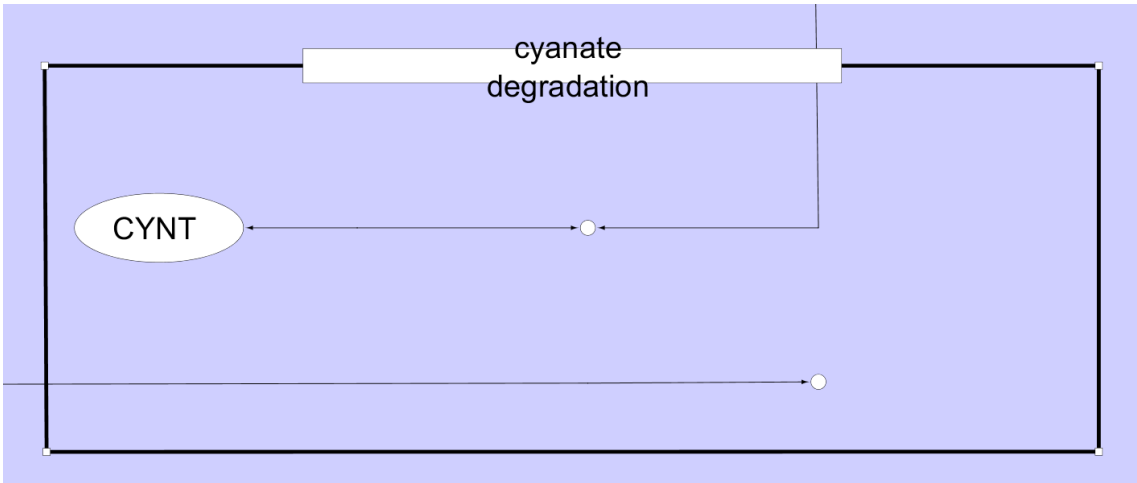


Abbildung 23 - Cyanat-Abbau aus der Stoffwechselkarte dargestellt im niedrigsten Detaillevel [Abbildung Cytoscape]

Im niedrigsten Detaillevel sind von jedem Stoffwechselweg nur noch die Zentralmetaboliten explizit dargestellt. Dies erleichtert die Übersicht über die Gesamtkarte, wenn beispielsweise nach einem Stoffwechselweg gesucht wird. Hier würden zu viele eingeblendete Informationen nur stören. Dargestellt ist dieses Detaillevel in Abbildung 23.

Mit Hilfe dieser vier Detaillevel ist eine übersichtliche und dennoch detailreiche Darstellung der generischen Stoffwechselkarte möglich.

3.4.2 Organismen-spezifische Ansicht

Auch die Ergebnisse des Programms EnzymeDetector können auf der Stoffwechselkarte dargestellt werden. Zunächst wird aus der entsprechenden Datenbank der Enzymvorrat des gewünschten Organismus abgefragt.

Anschließend werden Teile der Stoffwechselwege in der Stoffwechselkarte ausgegraut, die laut Enzymvorrat des Organismus nicht beschritten werden können. Somit ist es möglich, auf einen Blick zu erkennen, welche Stoffwechselwege abgedeckt sind bzw. zu welchen Teilen.

In Abbildung 24 ist ein Beispiel für *C. glutamicum* abgebildet. Alle Enzyme des Coenzym-A-Metabolismus sind in diesem Organismus vorhanden, somit sind auch alle farbig dargestellt. Man sieht auf einen Blick, dass dieser Stoffwechselweg keine Lücken aufweist.

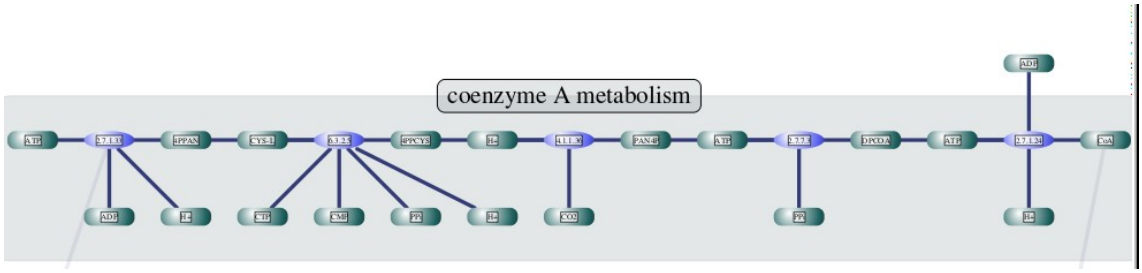


Abbildung 24 - Darstellung des Coenzym-A-Metabolismus aus C. glutamicum [Abbildung BRIME]

Im Gegensatz dazu zeigt Abbildung 25 einen Weg, der nicht vollständig abgedeckt ist. Dargestellt ist der Citratzyklus von *C. glutamicum*. Es ist klar erkennbar, dass einzelne Enzyme fehlen. Beispielsweise die Succinatdehydrogenase. Allerdings ist auch auf einen Blick zu sehen, dass es eine Umgehung des Enzyms im Organismus gibt.

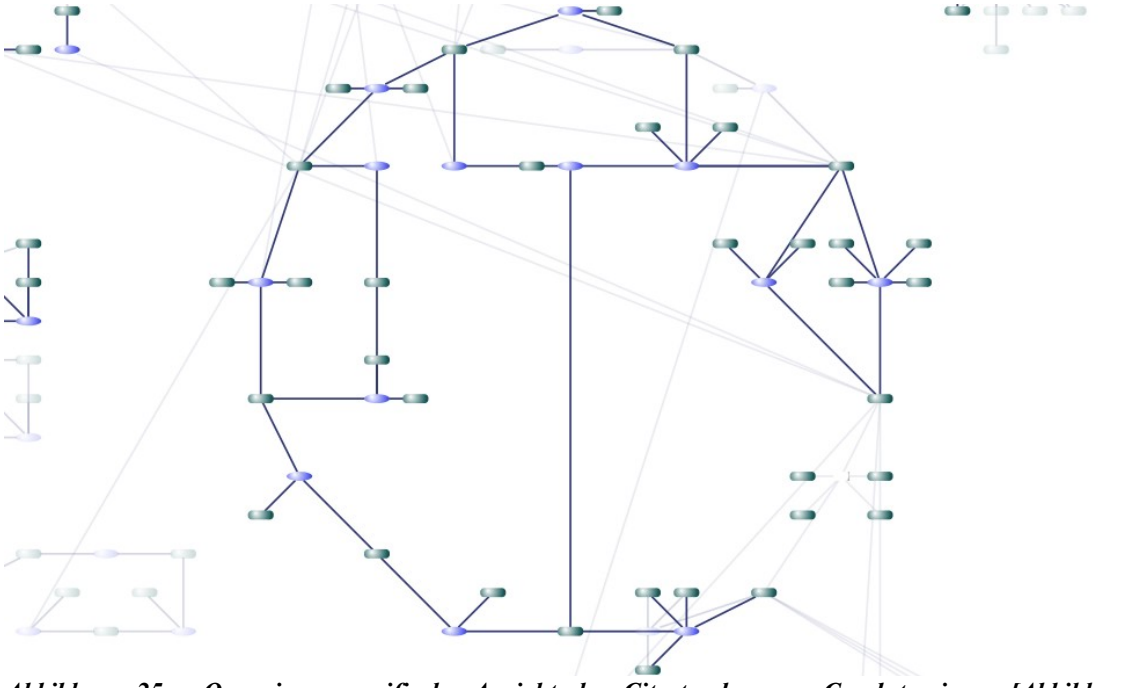


Abbildung 25 - Organismenspezifische Ansicht des Citratzyclus aus C. glutamicum [Abbildung BRIME]

Die organismenspezifische Ansicht ermöglicht es dem Nutzer, einen schnellen Überblick über den Metabolismus eines Organismus zu erhalten. Es ist leicht zu

erkennen, welche wichtigen Enzyme bis jetzt in der Genomannotation fehlen, um erforderliche Stoffwechselwege komplett abzudecken. Mit dieser Information kann einerseits gezielt nach diesen Enzymen in den EnzymeDetector Ergebnissen gesucht werden, da es möglich ist, dass sie nur auf Grund des gewählten Grenzwertes nicht in den Enzymvorrat des Organismus übernommen wurden. Andererseits kann diese Information Hinweise für die Laborarbeit geben, nach welcher Enzymaktivität gezielt untersucht werden sollte.

3.4.3 Visualisierung von Flussverteilungen

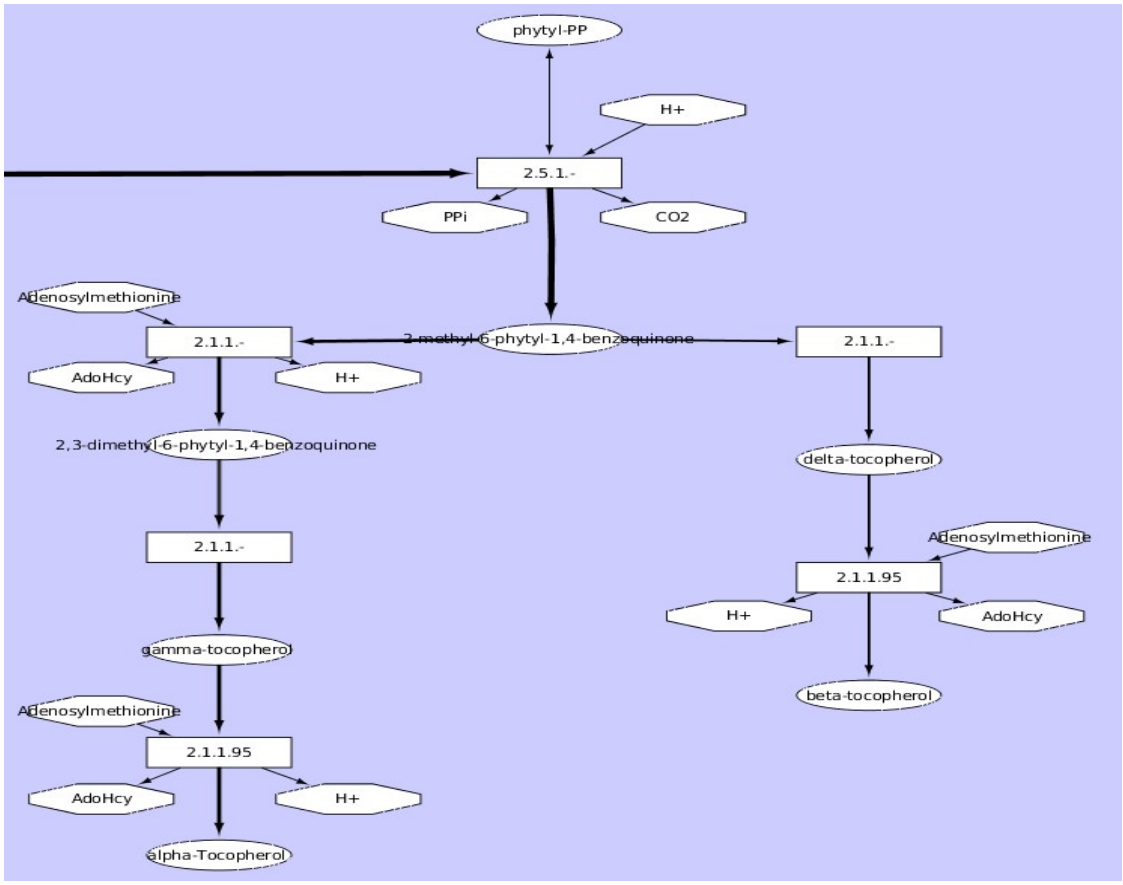


Abbildung 26 - Darstellung der Flüsse einer FBA am Beispiel des Vitamin E-Metabolismus [Abbildung Cytoscape]

Eine weitere Möglichkeit der Kartenansicht ist das Einbinden von Flussinformationen, zum Beispiel aus einer Flux Balance Analyse (FBA). Voraussetzung für diese Möglichkeit ist die einheitliche Vergabe von Reaktions-Identifiern. Für die Fälle, wo eine solche einheitliche Vergabe nicht vorlag, wurde ein Programm für den Abgleich der

im metabolischen Modell genutzten Reaktionen mit den Reaktionen aus der Karte implementiert[57].

Dargestellt werden die Beträge der Flüsse mittels der Dicke der Kanten. Die Richtungen der Flüsse werden mit Pfeilen dargestellt. Dies ist in Abbildung 26 an einem Beispiel veranschaulicht. Je dicker eine Kante ist, desto höher ist die Reaktionsgeschwindigkeit mit der die Stoffwechselprodukte durch diese Reaktion fließen.

Mit Hilfe dieser Ansicht kann schnell ermittelt werden, welche Teile des Netzwerkes bei den vorliegenden Modellparametern aktiv sind. So lassen sich die einzelnen Wege in ihrem metabolischen Kontext wesentlich leichter erfassen. Auch lassen sich durch die grafische Darstellung die Ergebnisse leichter auswerten. Die Suche nach essentiellen Stoffwechsel-Hubs im Netzwerk wird erleichtert.

3.5 AtomMapper

Im Rahmen der vorliegen Arbeit wurde ein bestehendes Atom-Zuordnungsprogramm (siehe2.17) erweitert. Der bestehende Algorithmus ermöglichte den Vergleich zweier Moleküle miteinander und lieferte als Ergebnis, welche Atome des einen Moleküls welchen Atomen des anderen Moleküls zugeordnet werden konnten. Das Programm heißt AtomMapper.

3.5.1 Beachtung der Bindungsart für bestimmte funktionelle Gruppen

Da im Grundalgorithmus nur die Atomsymbole, nicht aber die Bindungsart beachtet wurde, kam es besonders bei festen funktionellen Gruppen, wie Phosphat- und Carboxygruppen, zu mehreren gleichwertigen Ergebnissen auf Grund der vorhandenen Spiegelebene innerhalb dieser Gruppen. Der Grundalgorithmus konnte nicht zwischen der falschen und der richtigen Zuordnung von einem doppelbindigen Sauerstoff der Phosphatgruppe zu einem einfach-gebundenen unterscheiden.

Durch eine Erweiterung des Grundalgorithmus werden nun für Carboxy- und Phosphatgruppen die Bindungsarten als Information hinzugenommen. Dies verkleinert den Lösungsraum für Phosphatgruppen von 3 möglichen Zuordnungen auf 2 und für Carboxygruppen von 2 Lösungen auf 1. Es werden jetzt nicht mehr einfach-gebundene Atome dieser funktionellen Gruppen auf zweifach-gebundene Atome zugeordnet.

Für Phosphatendgruppen gibt es aber weiterhin immer zwei gleichberechtigte Lösungen, bei denen die zwei einfach-gebundenen Sauerstoff-Atome jeweils einander zugeordnet werden.

3.5.2 Einlesen ganzer Reaktionen

Der Grundalgorithmus erlaubte nur den Vergleich von zwei Molekülen miteinander. Für die vorliegende Arbeit wurde dieser Algorithmus in ein neues Programm, den AtomMapper, eingebettet, mit dem der Atomverlauf ganzer Reaktionen verfolgt werden kann. Die Stöchiometrie der Reaktionen wird berücksichtigt. Hierfür werden Moleküle,

die in einer Reaktion mehrfach vorkommen, einfach mehrfach in den übergebenen Molekülen aufgelistet.

Das neue Programm führt zunächst ein Preprocessing der Moleküle durch. Anschließend wird der eigentlichen Zuordnungsversuch mit einer abschließenden Prüfung der erhaltenen Lösung durchgeführt.

3.5.2.1 *Preprocessing der Moleküle*

Im Preprocessing werden die Moleküle nach Edukt- und Produktseite unterschieden und nach Größe sortiert. Diese Maßnahme sorgt für eine kürzere Rechenzeit des AtomMapper, da so zunächst die großen Moleküle beider Seiten in den Vergleich gehen. Wenn ein sehr kleines Molekül mit einem sehr großen Molekül verglichen wird, kann es sehr viele Möglichkeiten für die Zuordnung geben und der Lösungsraum ist sehr groß. Werden aber zunächst die großen Moleküle daraufhin geprüft, ob ihre Atome einander zugeordnet werden können, werden meist im ersten Schritt große Teile dieser beiden Moleküle von weiteren Zuordnungen ausgeschlossen und der Lösungsraum wird nicht unnötig groß.

Des Weiteren werden Wasserstoffatome vom Zuordnungsprozess ausgeschlossen, da sie in Molfiles oft nicht angegeben werden und außerdem der bestehende Grundalgorithmus diese aussortiert, falls sie in den Molfiles notiert sein sollten. Dieser Schritt ist wichtig, da am Ende des Algorithmus geprüft wird, ob alle Atome beider Reaktionsseiten zugeordnet werden konnten, wobei Wasserstoffatome immer übrig bleiben würden.

3.5.2.2 *Zuordnung der Atome*

Nach dem Preprocessing der Moleküle findet das eigentliche Mapping der Reaktionspartner statt. Damit alle möglichen Mapping-Kombinationen getestet werden, wurde der Algorithmus als Tiefensuche implementiert. Man kann sich die möglichen Kombinationen zwischen den Molekülen als Baum vorstellen, wobei jede Kombination einen Knoten des Baums bildet. Für die Tiefensuche wird ein Zweig dieses Baums durchlaufen, bis das Ende des Zweiges erreicht ist. Anschließend wird schrittweise zurückgegangen, bis ein bisher noch nicht besuchter Zweig des Graphen erreicht ist.

Dieser wird wieder bis zum Ende des Zweiges durchlaufen. Es handelt sich also um ein Backtracking-Verfahren, in dessen Rahmen alle Molekül-Kombinationen untersucht werden. Abbildung 27 verdeutlicht die Reihenfolge dieser Pre-Order-Traversierung.

Für jede Molekül-Paarung, also jeden Knoten des Kombinationsbaums, wird ein zweiteiliger Algorithmus ausgeführt. Dieser wird rekursiv aufgerufen, bis alle Moleküle der beiden Reaktionsseiten durchlaufen sind.

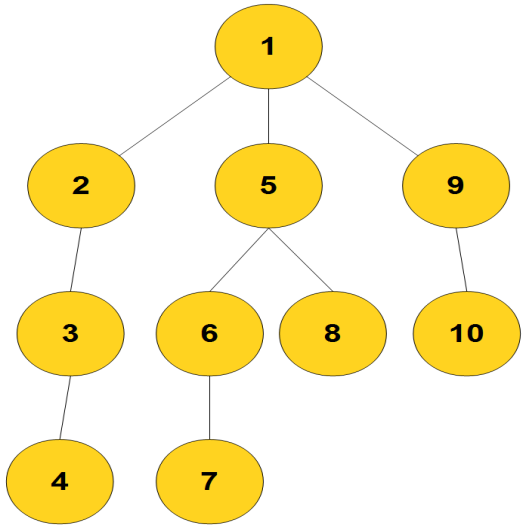


Abbildung 27 - Reihenfolge des Ablaufens der Knoten bei einer Tiefensuche (pre-order/depth-first)

Im ersten Schritt wird der Grundalgorithmus, der schon bestand, genutzt, um nach einer Atomzuordnung zwischen der aktuellen Molekülkombination zu suchen. Im nächsten Schritt wird geprüft, welche Atome von dieser Zuordnung bisher nicht betroffen sind, d.h. welche noch verfügbar sind. Hierbei wird unterschieden, ob es sich um mehrere zusammenhängende Atome handelt oder um Einzelatome.

Erstere werden als Pseudomolekül wieder in den Algorithmus eingespeist und stehen für weitere Zuordnungen zur Verfügung. Das bedeutet, dass dem Kombinationsbaum neue Knoten hinzugefügt werden. Die Einzelatome werden abgelegt und an späterer Stelle beim Erreichen des Zweigendes beachtet.

Ist das Ende eines Zweiges des Kombinationsbaums erreicht, wird die Lösung auf Zulässigkeit geprüft. Diese Prüfung erfolgt nach drei Kriterien:

1. Ist die aktuell gefundene Lösung besser oder gleichwertig zu der bis jetzt besten Lösung?

2. Wurden alle Moleküle oder Pseudomoleküle der Reaktion einander zugeordnet?

3. Können die gespeicherten Einzelatome der beiden Reaktionsseiten einander zugeordnet werden und zwar ohne, dass Atome übrig bleiben?

Nur wenn alle drei Fragen mit ja beantwortet werden können, wird die Lösung gespeichert. Ist die Lösung besser als eine bisher gefundene, wird diese überspeichert, ansonsten wird die Lösung als gleichwertig hinzugefügt.

Da es manchmal Fehler in den Molfiles gibt, wird die Lösung auch dann abgelegt, wenn nur die ersten beiden Fragen mit ja beantwortet werden konnten. Findet sich keine Lösung, für die alle Fragen mit ja beantwortet werden konnten, wird die Lösung ausgegeben, bei der eine erfolgreiche Zuordnung der Einzelatome nicht möglich war.

3.5.3 Ausgabe des AtomMappers

Das Programm gibt sein Ergebnis entweder als Text aus oder in einem verschachtelten Vektor. Dieser Vektor kann, wenn die entsprechenden Klassen des AtomMappers eingebunden wurden, aus Fremdprogrammen aufgerufen werden, und das Ergebnis kann zur Weiterberechnung dienen.

Dies geschieht im Rahmen des Programms gapFiller. Hier wird (u.a.) auf Grundlage der Atomzuordnungen entschieden, welche Wege in einem Stoffwechselnetzwerk von einem Molekül zu einem anderen genommen werden dürfen. Dadurch werden Wege ausgeschlossen, bei denen keine Atome des Ausgangsmetaboliten mehr im Endmetaboliten vorhanden sind.

In Abbildung 28 ist ein Beispiel für die Textausgabe des Programms dargestellt. Es werden zuerst die beiden zugeordneten Reaktionspartner bzw. ihre Compound-ID[2]

2 Wird aus dem Namen der übergebenen Molfiles ermittelt, wenn möglich.

angegeben und dann in einer Liste die Nummern der Atome der beiden Moleküle, die einander zugeordnet werden konnten.

```
###Mappings zwischen folgenden Molekuelen:###
Molfiles/C00817.mol  Molfiles/C00204.mol
final mapping Nr. 1   Laenge: 12
1 6
2 4
5 3
9 2
13 1
3 8
7 9
11 11
6 7
8 10
10 5
12 12

###Einzelatommappings###
Varianten des Einzelatommappings:
Molekuel: Molfiles/C00817.mol Atom: 4        Molekuel: Molfiles/C00001.mol Atom: 1
```

Abbildung 28 - Ausgabe des Programms AtomMapper für die Zuordnung der Reaktion von D-Altronate (C00817) zu 2-Dehydro-3-deoxy-D-gluconate (C00204) unter Wasserabspaltung (C00001)

Links stehen die Atome des Reaktionspartners der Eduktseite, rechts die des Reaktionspartners der Produktseite. Die Zahlen der Atome beziehen sich auf die Atomzahlen entsprechend der Molfiles. Da im Algorithmus keine Wasserstoff-Atome beachtet werden, bleibt vom Wassermolekül nur das Sauerstoff-Atom übrig. Auf Grund dessen wird das gesamte Molekül als Einzelatom betrachtet und taucht somit auch erst unter den Einzelatom-Zuordnungen auf. Diese werden im Anschluss an die Atomzuordnungen der verschiedenen Molekülkombinationen ausgegeben.Zur Veranschaulichung der Ergebnisse können Programme wie MarvinView[58] zu Hilfe genommen werden. Dieses Programm ermöglicht eine Ansicht der Molfiles als Strukturformel, bei der die zugehörigen Atomzahlen eingeblendet werden können. Dies erleichtert die manuelle Auswertung der Atomzuordnungen durch das Programm.

3.5.4 Fallbeispiele der Atom-Zuordnung

3.5.4.1 *Einfache Gruppenübertragung*

In Abbildung 29 ist die Reaktion von Fructose-6-phosphat zu Fructose-1,6-bisphosphat zu sehen, wobei ADP zu AMP und einem Phosphatrest gespalten wird. Katalysiert wird diese Reaktion vom der ADP-abhängige Phosphofructokinase. Es handelt sich um eine einfache Reaktion mit wenigen Reaktionspartnern. Die Atomzuordnung dieser Reaktion wird vom Algorithmus innerhalb von Millisekunden erstellt.

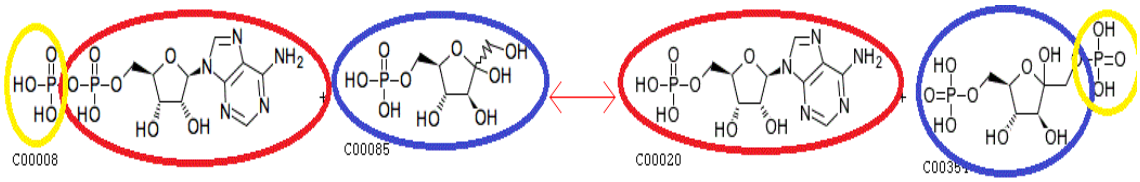


Abbildung 29 - Reaktion von Fructose-6-phosphat zu Fructose-1,6-bisphosphat und Zuordnung der einzelnen Gruppen, deren Atome einander zugeordnet werden können

Der Algorithmus ermittelt die einzelnen Atomgruppen, die einander zugeordnet werden können. Diese sind in Abbildung 29 jeweils farblich passend auf den beiden Reaktionsseiten markiert. Da in jeder der Gruppen eine Phosphat-Endgruppe vorkommt, gibt es zwar nur eine Möglichkeit der Gruppenzuordnung, aber für jede Zuordnung der Gruppen zueinander zwei gleichwertige Möglichkeiten. Dies liegt daran, dass der Algorithmus nicht zwischen den zwei einfach-gebundenen Sauerstoff-Atomen unterscheiden kann.

Viele Reaktionen laufen nach genau dem gleichen Prinzip ab, wie bei diesem Beispiel zu sehen ist. Die Gruppe eines Moleküls wird auf ein anderes Molekül übertragen.

3.5.4.2 *Reaktion mit Molekülen mit Spiegelebene*

In Abbildung 30 ist ein weiteres Beispiel für ein Ergebnis aus dem AtomMapper zu sehen. Es handelt sich um die hydrolytische Spaltung von Digallat zu zwei Methylgallat-Molekülen. Da bei dieser Reaktion zweimal das gleiche Molekül entsteht, gibt es zwei gleichwertige Lösungen für das Zuordnen der Atome. Diese sind in der Abbildung unter A und B dargestellt.

Des Weiteren kann das Sauerstoffatom aus dem Wassermolekül zwei Atomen aus dem Methylgallat zugeordnet werden, da dieses für das Ringsystem spiegelgleich ist. Diese Tatsache wird in der Abbildung dadurch veranschaulicht, dass die zweite Atomzuordnungsmöglichkeit mit gestrichelter Umrandung dargestellt ist.

Auch bei der rot markierten Reaktionsgruppe des Digallat gibt es zwei Sauerstoffatome, die der AtomMapper nicht unterscheiden kann. Also gibt es für diese Gruppenzuordnung ebenfalls zwei Möglichkeiten. Die Sauerstoffatome sind zwar einmal einfach- und einmal doppelt-gebunden, die Bindungsart untersucht der Algorithmus allerdings nur für Carboxy- und Phosphatgruppen.

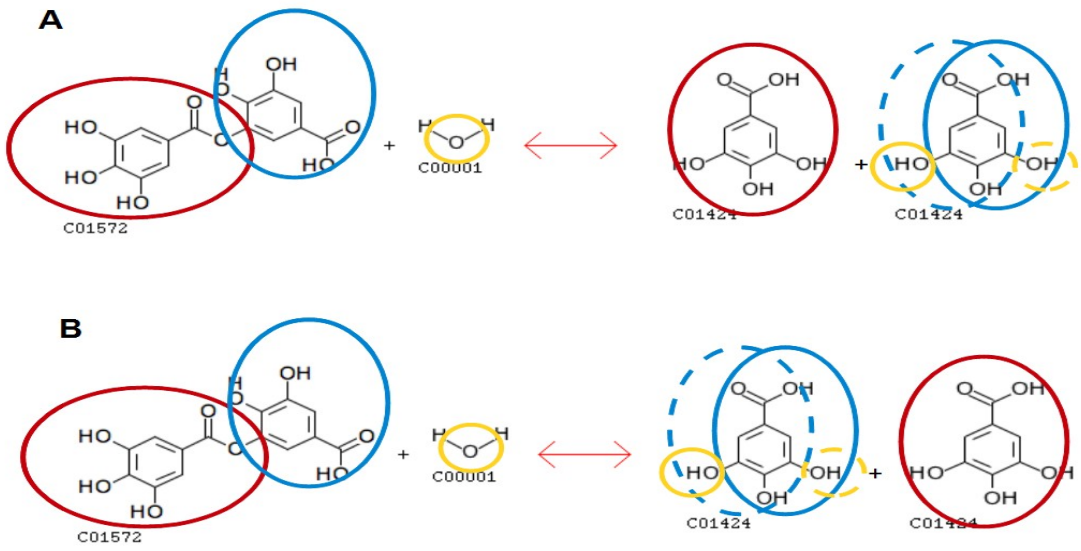


Abbildung 30 - Beide Möglichkeiten der Zuordnung der Atomgruppen, wie sie durch den AtomMapper gefunden wurden, der hydrolytischen Spaltung von Digallat

Die Ausgabe des AtomMapper zu der in Abbildung 30 dargestellten Reaktion ist im Anhang zu finden.

Das zweite Fallbeispiel zeigt, dass selbst bei recht einfachen Reaktionsmechanismen der Lösungsraum des AtomMapper recht groß sein kann. Dies tritt besonders auf, wenn einer oder mehrere der Reaktionspartner eine Spiegelebene innerhalb des Moleküls haben. Da in diesem Fall die Atome nicht unterscheidbar sind, muss der AtomMapper hier zwei gleichberechtigte Lösungen liefern.

3.6 Fazit und Ausblick

Die vorliegende Arbeit beleuchtet einen Teil der Probleme, die auf dem Weg vom Genom zum Modell entstehen, und bietet durch die vorgestellten Programme Hilfen und Lösungen.

Wie gezeigt wurde, ist es nötig, die Genomannotationen mehrerer Quellen zu integrieren, um einen umfassenden Einblick in die metabolischen Möglichkeiten eines Organismus zu erhalten. Besonders die BLAST-basierte Genomannotation liefert hier einen Zugewinn an Information. Auch wenn manche dieser gefundenen Annotationen eine nicht so hohe Güte haben, haben immerhin 1/3 der BLAST-Annotationen einen E-value von unter 10^{-120}. Neben der Tatsache dass nur durch die integrierte Annotation eine hohe Informationsdichte erreicht werden kann, bietet nur das Einbeziehen eines Wahrheitsmaßes und der Vergleich der Informationen der verschiedenen Quellen die Möglichkeit, die Güte der gefundenen Annotationskandidaten zu beurteilen. Dies ist dringend erforderlich, da die Annotationen der verschiedenen Quellen nur in etwa einem Viertel der Fälle übereinstimmen.

Zur weiteren Verbesserung der Informationsabdeckung und der Güte der Daten des EnzymeDetectors ist es sinnvoll, weitere Quellen zu integrieren. Besonders vorteilhaft wäre es, Quellen hinzuzufügen, die nicht auf Sequenzähnlichkeiten basieren, da die bisher eingebundenen Quellen dies tun. Generell sind alle Quellen geeignet, die Genomannotationen für ein breites Feld an Organismen anbieten und die ein automatisches Herunterladen der Informationen ermöglichen. Alternativ könnten Informationen aus Programmen eingebunden werden, die automatisch eine Genomannotation erstellen.

Die im Rahmen des EnzymeDetector durchgeführte Operonanalyse ist zwar nicht signifikant genug, um eine eigenständige Genomannotation zu liefern, allerdings konnte für ein Zehntel der Fälle, in denen die BLAST-Suche bislang keine klare Entscheidung ergab, mit den Ergebnissen der Analyse der Lösungsraum eingeschränkt werden.

Zukünftig sollte versucht werden, die Operonanalyse soweit zu verbessern, dass die Ergebnisse aussagekräftig genug für eine unabhängige Genomannotation sind.

Die Darstellung der Ergebnisse des EnzymeDetector auf einem Webinterface ermöglicht einen leichten Zugriff auf die Daten.

Auch wenn das Webinterface schon viele Funktionalitäten und Auswertungen anbietet, sollte dieses Angebot in Zukunft noch erweitert werden. Besonders interessant wäre hier die Möglichkeit, die Daten auch visuell darzustellen, wie es mit Hilfe der Stoffwechselkarte möglich ist.

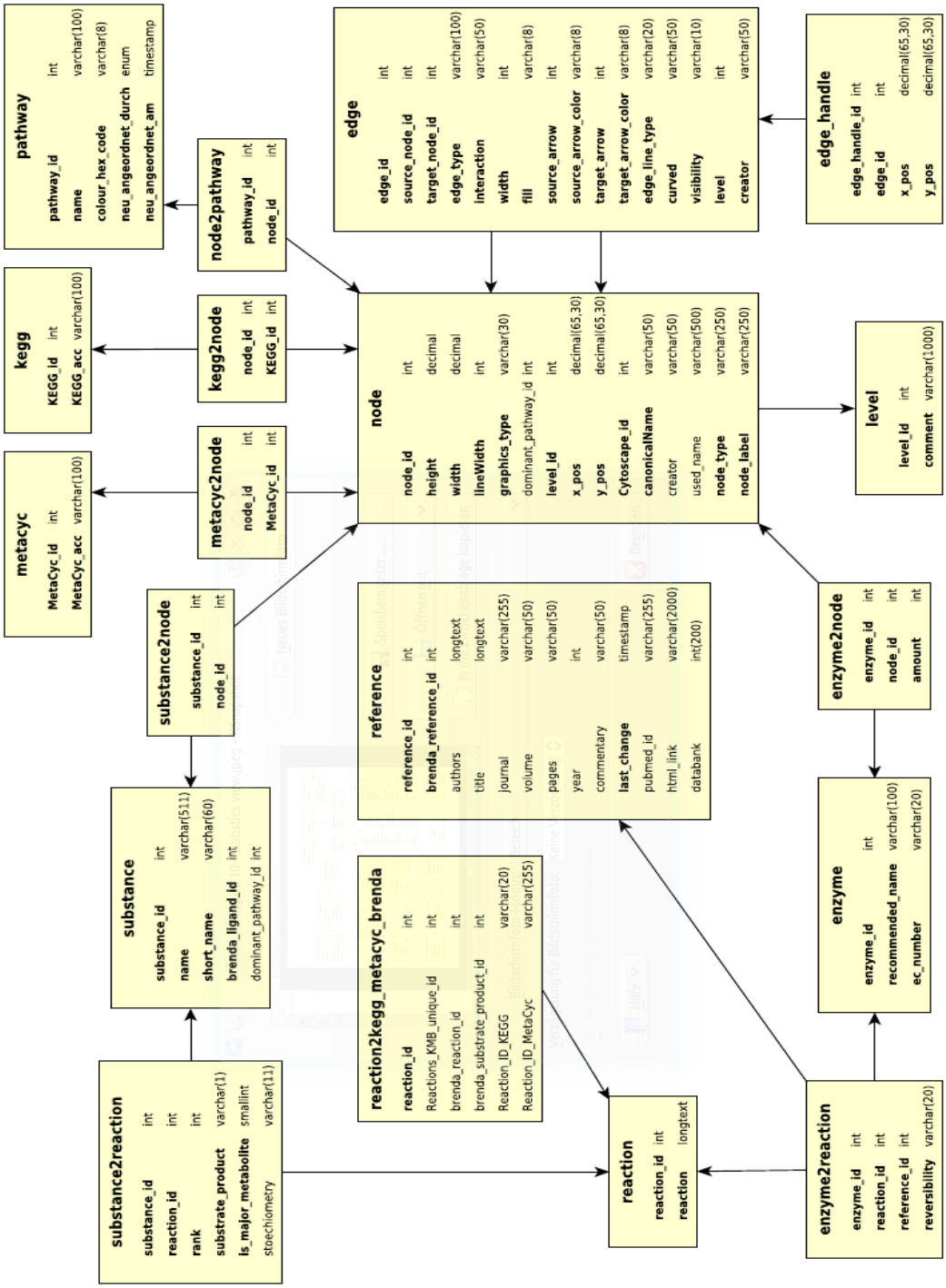
Die erstellte Stoffwechselkarte bietet die Möglichkeit, die Stoffwechselwege in verschiedenen Detailstufen zu betrachten und ermöglicht auch die Darstellung von Flüssen aus einer Flux-Balance-Analyse. So erhält der Nutzer einen schnelleren Überblick, welche Stoffwechselwege in einem Organismus laut Annotation vorhanden sind und wo noch Lücken im Netzwerk bestehen.

Die Daten dieser Karte werden laufend erweitert und verbessert. Besonders wichtig ist hier die Gewährleistung der Anbindung an andere Informationsquellen, zum Beispiel die Zuordnenbarkeit der Reaktionen zu den Reaktionen, die in Datenbanken wie BRENDA, Metayc und KEGG hinterlegt sind.

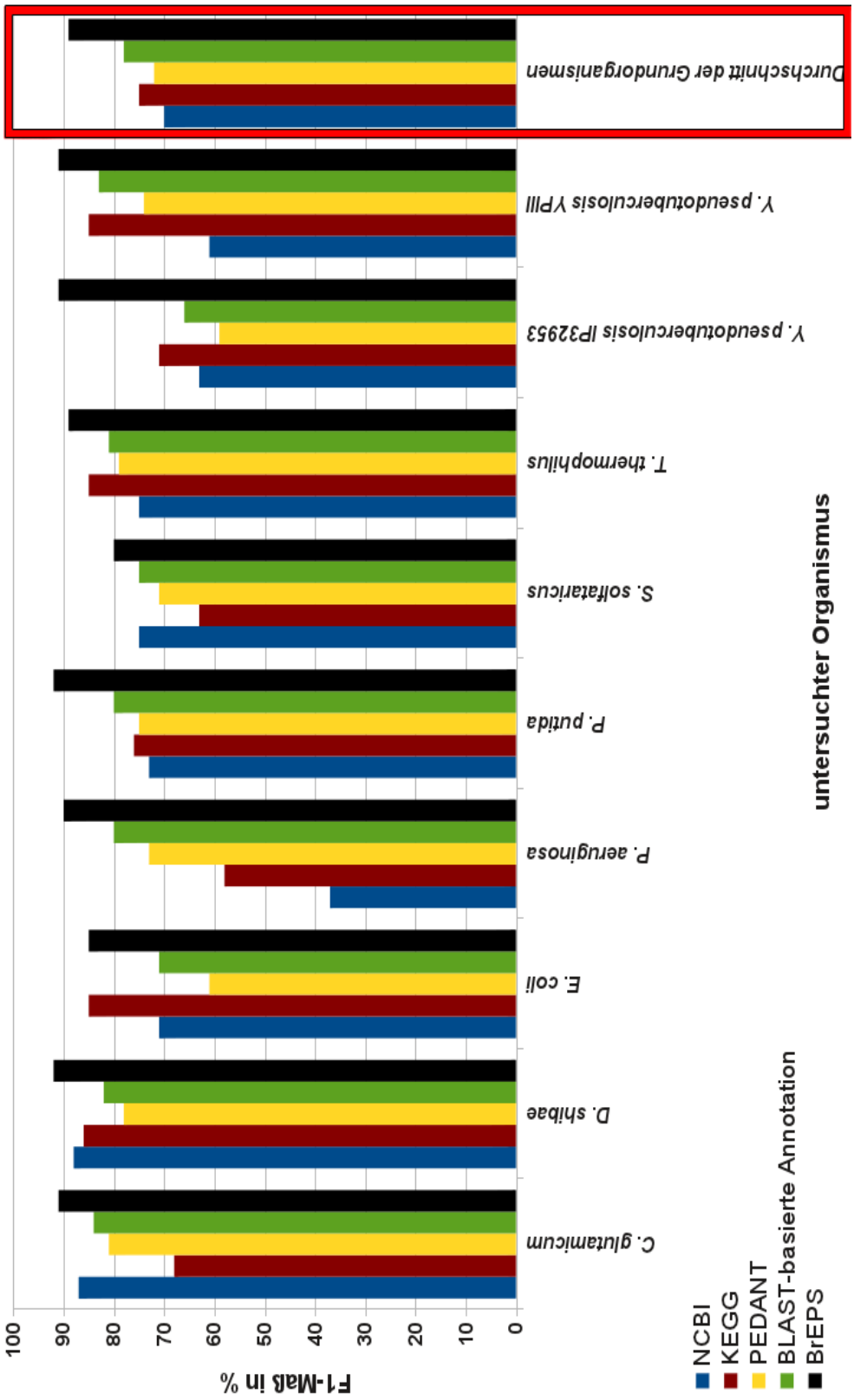
Durch den Zusammenschluss der vorgestellten Programme, wird das Erstellen von Modellen basierend auf der Genomannotation erleichtert. Es können Schritte automatisiert ausgeführt werden, die vorher manuell und zeitaufwendig erledigt werden mussten. Auch wenn weiterhin noch eine manuelle Erweiterung der Genomannotation vonnöten ist, so bieten die Programme doch eine große Hilfe und Zeitersparnis.

Anhang

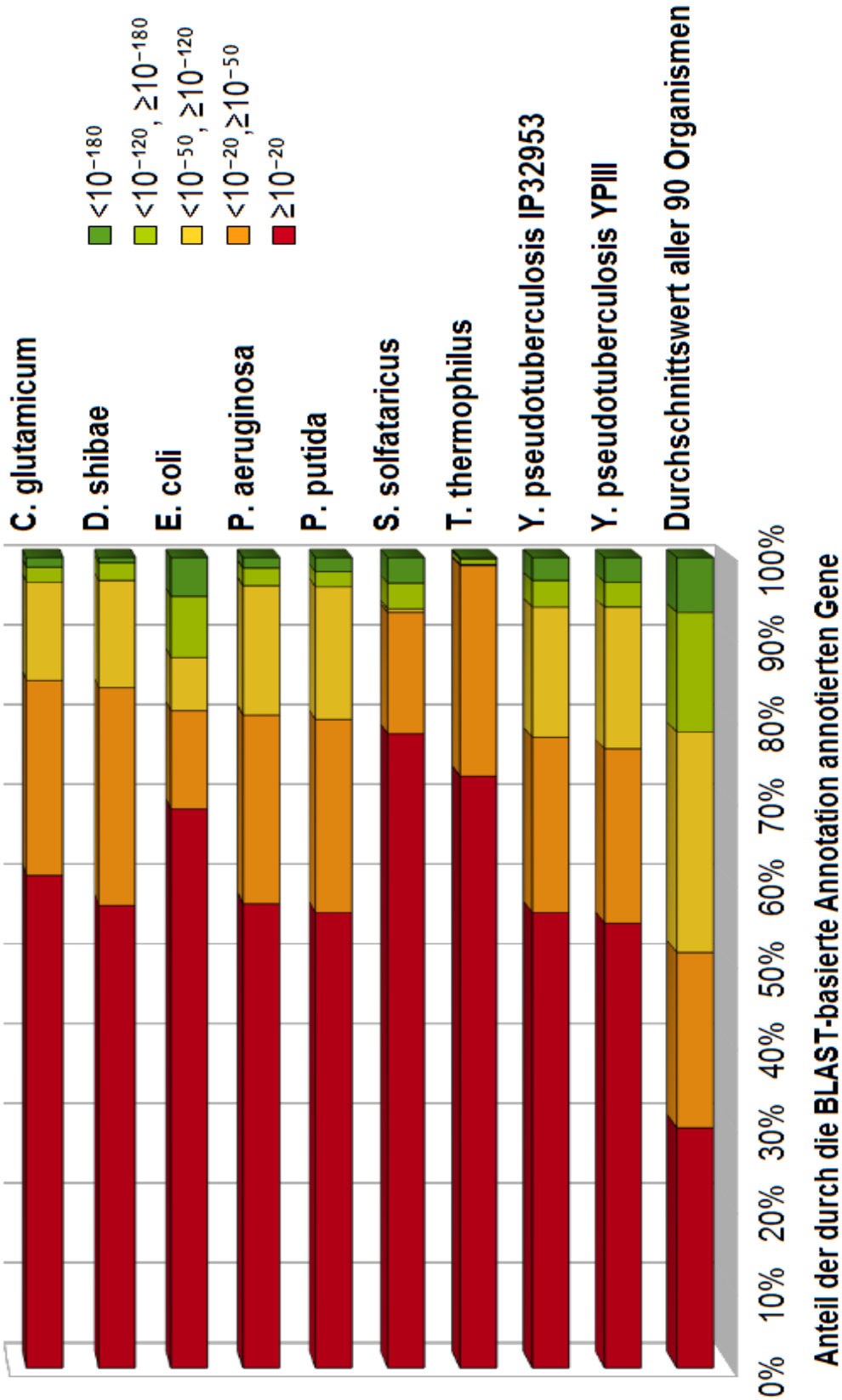
Anhang 1 – Tabellen der SQL-Datenbank „metabolic_pathways“ und ihre Beziehungen untereinander (Abbildung aus [57] (Seite 19)



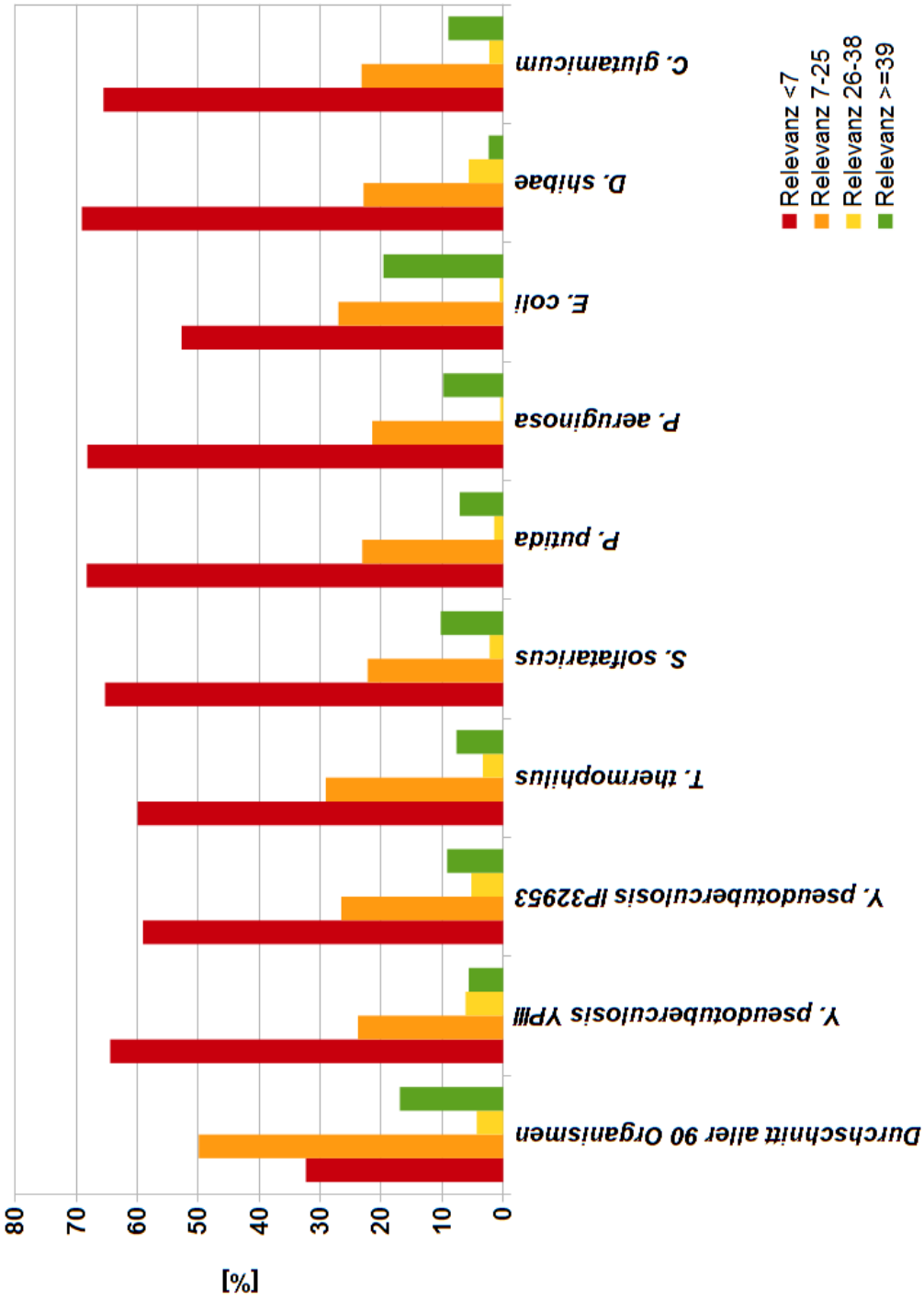
Anhang 2 – F1-Maß der einzelnen Annotationsquellen der Referenzorganismen und der Durchschnittswerte dieser neun Organismen (Seite48)



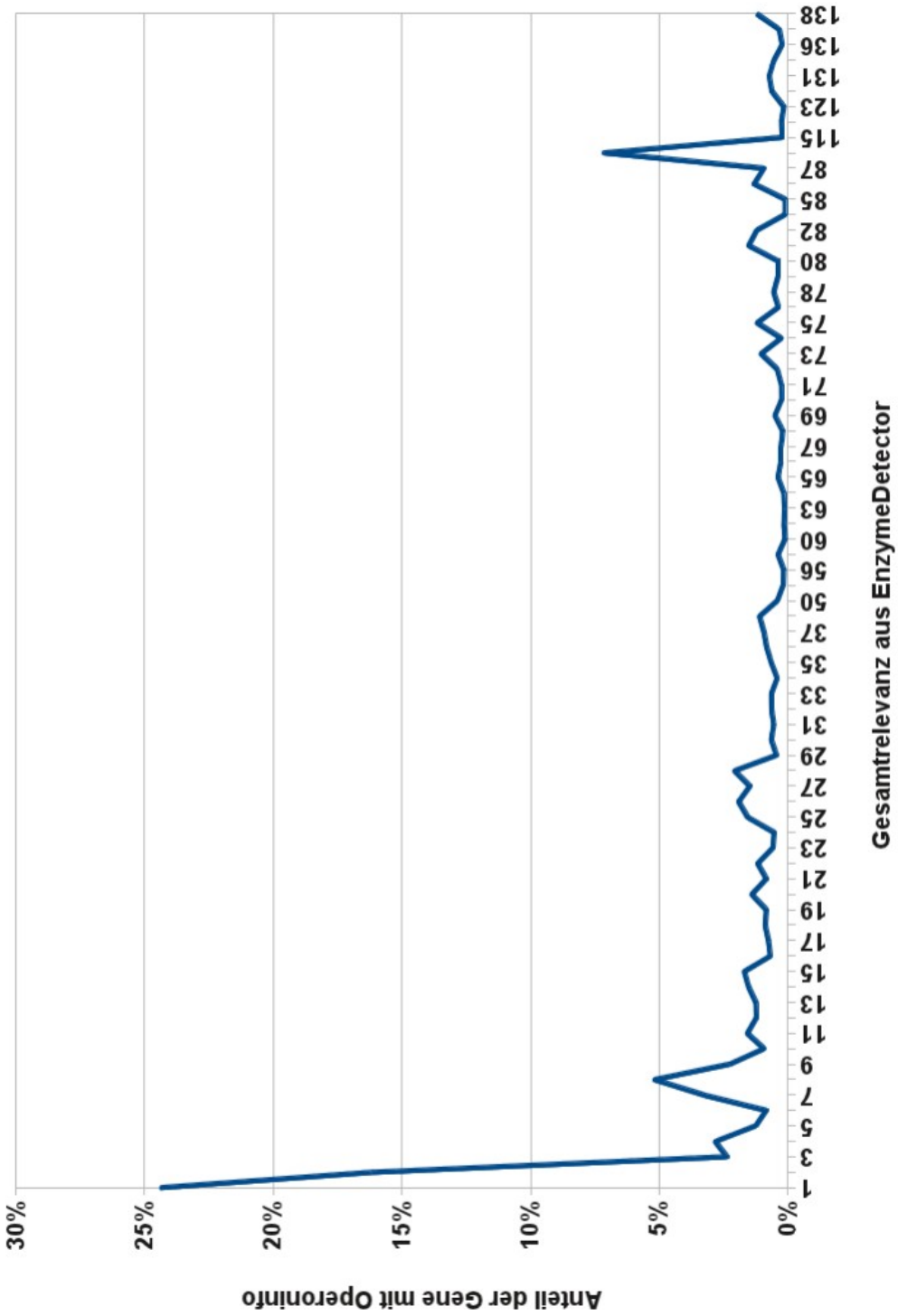
Anhang 3 – E-value Güteklassen der Annotationen aus der BLAST-basierten genomannotation (Seite 63)



Anhang 4 – Relevanzgruppen der EnzymeDetector Ergebnisse der neun Referenzorganismen und der Durchschnitt dieser Werte über alle 90 Organismen (Seite 65)



Anhang 5 – Daten aus der Operonanalyse aufgetragen gegen die Relevanz mit der diese Annotationen im EnzymeDetector gefunden wurden (Seite 69)



Anhang 6 – Liste der Organismen, die inklusive der BLAST-basierten Genomannotation vom EnzymeDetector berechnet wurden

Corynebacterium_glutamicum	ATCC_13032
Dinoroseobacter_shibae	DFL_12
Escherichia_coli	K_12_MG1655
Pseudomonas_aeruginosa	PA01
Pseudomonas_putida	KT2440
Sulfolobus_solfataricus	P2
Thermus_thermophilus	HB27
Yersinia_pseudotuberculosis	IP32953
Yersinia_pseudotuberculosis	YPIII
Escherichia_coli	O157:H7_Sakai
Sulfolobus_tokodaii	strain7
Yersinia_pestis	CO92
Escherichia_coli	K_12_W3110
Corynebacterium_glutamicum	ATCC_13032
Escherichia_coli	O6:K2:H1_CFT073
Pseudomonas_syringae	pv_tomato_DC3000
Corynebacterium_diphtheriae	gravis_NCTC13129
Rhodopseudomonas_palustris	CGA009
Yersinia_pestis	91001
Thermus_thermophilus	HB8
Pseudomonas_syringae	pv_syringae_B728a
Pseudomonas_fluorescens	Pf_5
Sulfolobus_acidocaldarius	DSM_639
Corynebacterium_jeikeium	K411
Pseudomonas_syringae	pv_phaseolicola_1448A
Rhodopseudomonas_palustris	HaA2
Rhodopseudomonas_palustris	BisB18
Escherichia_coli	O18:K1:H7_UTI89
Rhodopseudomonas_palustris	BisB5
Pseudomonas_entomophila	L48
Yersinia_pestis	Antiqua
Yersinia_pestis	Nepal516
Escherichia_coli	O6:K15:H31_536
Rhodopseudomonas_palustris	BisA53
Pseudomonas_aeruginosa	UCBPP_PA14
Escherichia_coli	O1:K1:H7
Hyperthermus_butylicus	DSM_5456
Yersinia_enterocolitica	subsp_enterocolitica_8081
Staphylothermus_marinus	F1
Corynebacterium_glutamicum	R
Yersinia_pestis	Pestoides_F
Pseudomonas_mendocina	ymp
Pseudomonas_stutzeri	A1501
Pseudomonas_putida	F1
Pseudomonas_aeruginosa	PA7
Yersinia_pseudotuberculosis	IP31758
Escherichia_coli	O139:H28_E24377A
Escherichia_coli	O9_HS
Yersinia_pestis	Angola
Pseudomonas_putida	GB_1
Escherichia_coli	K_12_DH10B
Escherichia_coli	SMS_3_5
Pseudomonas_putida	W619
Corynebacterium_urealyticum	DSM_7109
Escherichia_coli	C_ATCC_8739
Rhodopseudomonas_palustris	TIE_1
Escherichia_coli	O157:H7_EC4115
Escherichia_coli	O152:H28_SE11
Escherichia_fergusonii	ATCC_35469
Pseudomonas_aeruginosa	LESB58
Escherichia_coli	55989
Escherichia_coli	O81_D1a
Escherichia_coli	O8_IAI1
Escherichia_coli	O7:K1_IAI39
Escherichia_coli	O45:K1:H7_S88
Escherichia_coli	O17:K52:H18_UMN026
Corynebacterium_aurimucosum	ATCC_700975
Sulfolobus_islandicus	LS215
Sulfolobus_islandicus	M1425
Pseudomonas_fluorescens	SBW25
Sulfolobus_islandicus	M1627
Sulfolobus_islandicus	YG5714
Sulfolobus_islandicus	YN1551
Sulfolobus_islandicus	M164
Escherichia_coli	BW2952
Escherichia_coli	BL21_Gold
Escherichia_coli	B_REL606
Escherichia_coli	O157:H7_TW14359
Escherichia_coli	O103:H2_12009
Escherichia_coli	O111:H__11128
Escherichia_coli	O26:H11_11368
Rhodothermus_marinus	DSM_4252
Sulfolobus_islandicus	LD85
Escherichia_coli	O55:H7_CB9615
Meiothermus_ruber	DSM_1279
Bacillus_megaterium	QM_B1551
Yersinia_pestis	Z176003
Bacillus_megaterium	DSM_319
Staphylothermus_hellenicus	DSM_12710
Syntrophothermus_lipocalidus	DSM_12680

Anhang 7 – Output des Programms AtomMapper für die hydrolytische Spaltung von Digallat

```
###Mappings zwischen folgenden
Molekuelen:###

/home/squ/workspace/Molfiles/C01572.mol
/home/squ/workspace/Molfiles/C01424.mol

final mapping Nr. 1   Laenge: 12
20 9
18 6
16 3
11 1
6 4
12 8
17 2
19 5
22 10
21 11
23 12
3 7

final mapping Nr. 2   Laenge: 12
20 9
18 6
16 3
11 1
6 4
12 7
17 2
19 5
22 10
21 11
23 12
3 8

final mapping Nr. 3   Laenge: 12
20 9
18 5
16 2
11 1
6 4
12 8
17 3
19 6
22 11
21 10
23 12
3 7

final mapping Nr. 4   Laenge: 12
20 9
18 5
16 2
11 1
6 4
12 7
17 3
19 6
22 11
21 10
23 12
3 8
```

```
###Mappings zwischen folgenden
Molekuelen:###

/home/squ/workspace/Molfiles/C01572.mol
/home/squ/workspace/Molfiles/C01424.mol

final mapping Nr. 1   Laenge: 11
1 6
2 3
5 1
9 2
7 5
4 9
8 12
10 4
14 7
13 10
15 8

final mapping Nr. 2   Laenge: 11
1 5
2 2
5 1
9 3
7 6
4 9
8 12
10 4
14 7
13 11
15 8


###Einzelatommappings###
Varianten des Einzelatommappings:
Molekuel:
/home/squ/workspace/Molfiles/C00001.mol
Atom: 1        Molekuel:
/home/squ/workspace/Molfiles/C01424.mol
Atom: 11
-----------------------------------------
Molekuel:
/home/squ/workspace/Molfiles/C00001.mol
Atom: 1        Molekuel:
/home/squ/workspace/Molfiles/C01424.mol
Atom: 10
Molekuel:
/home/squ/workspace/Molfiles/C00001.mol
Atom: 1        Molekuel:
/home/squ/workspace/Molfiles/C01424.mol
Atom: 10
-----------------------------------------
```

```
###Mappings zwischen folgenden
Molekuelen:###

/home/squ/workspace/Molfiles/C01572.mol
/home/squ/workspace/Molfiles/C01424.mol
```

```
final mapping Nr. 1   Laenge: 12
20 9
18 6
16 3
11 1
6 4
12 8
17 2
19 5
22 10
21 11
23 12
3 7

final mapping Nr. 2   Laenge: 12
20 9
18 6
16 3
11 1
6 4
12 7
17 2
19 5
22 10
21 11
23 12
3 8

final mapping Nr. 3   Laenge: 12
20 9
18 5
16 2
11 1
6 4
12 8
17 3
19 6
22 11
21 10
23 12
3 7

final mapping Nr. 4   Laenge: 12
20 9
18 5
16 2
11 1
6 4
12 7
17 3
19 6
22 11
21 10
23 12
3 8


###Mappings zwischen folgenden
Molekuelen:###

/home/squ/workspace/Molfiles/C01572.mol
/home/squ/workspace/Molfiles/C01424.mol

final mapping Nr. 1   Laenge: 11
1 6
2 3
5 1
9 2
7 5
4 9
8 12
10 4
14 7
13 10
15 8

final mapping Nr. 2   Laenge: 11
1 5
2 2
5 1
9 3
7 6
4 9
8 12
10 4
14 7
13 11
15 8


###Einzelatommappings###
Varianten des Einzelatommappings:
Molekuel:
/home/squ/workspace/Molfiles/C00001.mol
Atom: 1       Molekuel:
/home/squ/workspace/Molfiles/C01424.mol
Atom: 11
----------------------------------------
Molekuel:
/home/squ/workspace/Molfiles/C00001.mol
Atom: 1       Molekuel:
/home/squ/workspace/Molfiles/C01424.mol
Atom: 10
Molekuel:
/home/squ/workspace/Molfiles/C00001.mol
Atom: 1       Molekuel:
/home/squ/workspace/Molfiles/C01424.mol
Atom: 10
```

Abkürzungsverzeichnis

AMENDA	Automatic Mining of ENzyme DAta
BLAST	Basic Local Alignment Search Tool
BRENDA	Braunschweig ENzyme DAtabase
BrEPS	Braunschweig Enzyme Pattern Search
BRIME	Braunschweig Interactive Metabolism Explorer
CSS	Cascading Style Sheets
DOM	Document Object Model
DOOR	Database of prOkaryotic OpeRons
E-value	Expect value
EC	Enzyme Commission
FBA	Flux Balance Analysis
FTP	File Transfer Protocol
GI	Gene Identifier
HTML	Hypertext Markup Language
KEGG	Kyoto Encyclopedia of Genes and Genomes
MCS	Maximal Common Subgraph
MOMA	Minimization of Metabolic Adjustment
NCBI	National Center for Biotechnology Information
PEDANT	Protein Extraction, Description and ANalysis Tool
RefSeq	Reference Sequence
SOAP	Simple Object Access Protocol
SQL	Structured Query Language
TrEMBL	Translated EMBL
UniParc	UniProt Archive
UniProt	Universal Protein Resource
UniProtKB	UniProt Knowledgebase
UniRef	UniProt Reference Clusters
XGMML	Extensible Graph Markup and Modeling Language

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

1. *Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW:* ***GenBank****. Nucleic Acids Res* 2011, **39**:D32-37.

2. **UniProt** [http://www.uniprot.org/].

3. Kitano H: **Perspectives on systems biology**. *New Gener Comput* 2000, **18**:199-216.

4. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**. *Nucleic Acids Res* 2007, **35**:D61-65.

5. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M: **KEGG for representation and analysis of molecular networks involving diseases and drugs**. *Nucleic Acids Res* 2010, **38**:D355-360.

6. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes**. *Nucleic Acids Res* 2000, **28**:27-30.

7. Riley ML, Schmidt T, Artamonova II, Wagner C, Volz A, Heumann K, Mewes H-W, Frishman D: **PEDANT genome database: 10 years online**. *Nucleic Acids Res* 2007, **35**:D354-357.

8. **The Universal Protein Resource (UniProt) in 2010**. *Nucleic Acids Res* 2010, **38**:D142-148.

9. Soh D, Dong D, Guo Y, Wong L: **Consistency, comprehensiveness, and compatibility of pathway databases**. *BMC Bioinformatics* 2010, **11**:449.

10. Poptsova MS, Gogarten JP: **Using comparative genome analysis to identify problems in annotated microbial genomes**. *Microbiology (Reading, Engl.)* 2010, **156**:1909-1917.

11. Furnham N, Garavelli JS, Apweiler R, Thornton JM: **Missing in action: enzyme functional annotations in biological databases**. *Nat. Chem. Biol* 2009, **5**:521-525.

12. Claudel-Renard C, Chevalet C, Faraut T, Kahn D: **Enzyme-specific profiles for genome annotation: PRIAM**. *Nucleic Acids Res* 2003, **31**:6633-6639.

13. Bairoch A: **The ENZYME database in 2000**. *Nucleic Acids Research* 2000, **28**:304-305.

14. Tian W, Arakaki AK, Skolnick J: **EFICAz: a comprehensive approach for accurate genome-scale enzyme function inference**. *Nucleic Acids Res* 2004, **32**:6226-6239.

15. Arakaki AK, Huang Y, Skolnick J: **EFICAz2: enzyme function inference by a combined approach enhanced by machine learning**. *BMC Bioinformatics* 2009, **10**:107.

16. Yang Y, Gilbert D, Kim S: **Annotation confidence score for genome annotation: a genome comparison approach**. *Bioinformatics* 2010, **26**:22-29.

17. Chitale M, Hawkins T, Park C, Kihara D: **ESG: extended similarity group method for automated protein function prediction**. *Bioinformatics* 2009, **25**:1739-1745.

18. Misra S, Harris N: **Using Apollo to browse and edit genome annotations**. *Curr Protoc Bioinformatics* 2006, **Chapter 9**:Unit 9.5.

19. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, Diekhans M, Dreszer TR, Giardine BM, Harte RA, Hillman-Jackson J, Hsu F, Kirkup V, Kuhn RM, Learned K, Li CH, Meyer LR, Pohl A, Raney BJ, Rosenbloom KR, Smith KE, Haussler D, Kent WJ: **The UCSC Genome Browser database: update 2011**. *Nucleic Acids Res* 2010.

20. Médigue C, Moszer I: **Annotation, comparison and databases for hundreds of bacterial genomes**. *Res. Microbiol* 2007, **158**:724-736.

21. Schnoes AM, Brown SD, Dodevski I, Babbitt PC: **Annotation error in public databases: misannotation of molecular function in enzyme superfamilies**. *PLoS Comput. Biol* 2009, **5**:e1000605.

22. Lang M, Schomburg: **A Tool for the Identification of Missing Enzymes in Biochemical Pathways**. In 2010.

23. Blum T, Kohlbacher O: **Using Atom Mapping Rules for an Improved Detection of Relevant Routes in Weighted Metabolic Networks**. *Journal of Computational Biology* 2008, **15**:565-576.

24. Akutsu T: **Efficient Extraction of Mapping Rules of Atoms from Enzymatic Reaction Data**. *Journal of Computational Biology* 2004, **11**:449-462.

25. Hattori M, Tanaka N, Kanehisa M, Goto S: **SIMCOMP/SUBCOMP: chemical structure search servers for network analyses**. *Nucleic Acids Research* 2010, **38**:W652-W656.

26. Leber M: **Koordinierung enzymatischer Reaktionen**. 2008.

27. Orth JD, Thiele I, Palsson BØ: **What is flux balance analysis?** *Nat Biotechnol* 2010, **28**:245-248.

28. **MetaCyc Encyclopedia of Metabolic Pathways** [http://metacyc.org/].

29. **IUBMB Nomenclature Home Page** [http://www.chem.qmul.ac.uk/iubmb/].

30. Alberts B, Bray D, Johnson A, Lewis J: **Genregulation**. In *Lehrbuch der molekularen Zellbiologie*. 2. Auflage. WILEY_VCH Verlag GmbH; :274 - 285.

31. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL: **BLAST+: architecture and applications**. *BMC Bioinformatics* 2009, **10**:421.

32. Merkl R: **Bioinformatik interaktiv**. In *Bioinformatik interaktiv : Grundlagen, Algorithmen, Anwendungen*. 2. Auflage. Weinheim: Wiley-VCH; 2009:159.

33. Bannert C, Welfle A, Aus dem Spring C, Schomburg D: **BrEPS: A flexible and automatic protocol to compute enzyme-specific sequence profiles for functional annotation**. *BMC Bioinformatics* 2010, **11**:589.

34. **MySQL :: MySQL 5.6 Reference Manual** [http://dev.mysql.com/doc/refman/5.6/en/index.html].

35. **KEGG - Current Statistics** [http://www.genome.jp/kegg/docs/statistics.html].

36. **NPO Bioinformatics Japan - KEGG FTP Academic Subscription** [http://www.bioinformatics.jp/en/keggftp.html].

37. **The NCBI Handbook - NCBI Bookshelf** [http://www.ncbi.nlm.nih.gov/books/NBK21101/].

38. **Index von ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/** [ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/].

39. **Index von ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/** [ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/].

40. Dam P, Olman V, Harris K, Su Z, Xu Y: **Operon prediction using both genome-specific and general genomic information**. *Nucleic Acids Res* 2007, **35**:288-298.

41. Mao F, Dam P, Chou J, Olman V, Xu Y: **DOOR: a database for prokaryotic operons**. *Nucleic Acids Res* 2009, **37**:D459-463.

42. **Operon Database at Computational Systems Biology Lab** [http://csbl1.bmb.uga.edu/OperonDB_10142009/DOOR.php].

43. **TAR-Verzeichnis DOOR** [https://csbl1.bmb.uga.edu/Operon_DB/10142009/opr1109.tar].

44. **Statistik NCBI** [ftp://ftp.ncbi.nih.gov/refseq/release/release-statistics/RefSeq-release46.03082011.stats.txt].

45. **Index von ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/** [ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/].

46. **KEGG: Kyoto Encyclopedia of Genes and Genomes** [http://www.genome.jp/kegg/].

47. **FTP-Server von KEGG** [ftp://ftp.genome.jp/pub/kegg/genes/organisms/].

48. **Pedant3 Databases** [http://pedant.gsf.de/webservices.jsp].

49. **PEDANT Webservice** [http://pedant.gsf.de/webservice].

50. **Index von ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/** [ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/].

51. **MySQLdb User's Guide** [http://mysql-python.sourceforge.net/MySQLdb.html].

52. **CTfile Formats**.

53. Lühr T, Scheer M, Busch M, Rother M, Schomburg D, Schomburg I: **BRIME A tool for visualization of metabolic models**. In 2011.

54. Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T: **Cytoscape 2.8: new features for data integration and network visualization**. *Bioinformatics* 2011, **27**:431-432.

55. **XGMML (eXtensible Graph Markup and Modeling Language)** [http://www.cs.rpi.edu/~puninj/xgmml.html].

56. **19.7. xml.dom.minidom — Lightweight DOM implementation — Python v2.7.2 documentation** [http://docs.python.org/library/xml.dom.minidom.html].

57. Busch M: **Verknüpfung der Reaktionen einer Stoffwechselkarte auf Datenbankebene und Visualisierung von metabolischen Flüssen im Kartenexport**. 2011.

58. **MarvinView « ChemAxon – toolkits and desktop applications for cheminformatics** [http://www.chemaxon.com/products/marvin/marvinview/].

Lebenslauf

Persönliche Daten

Name:	Susanne Quester
Geburtstag:	09.06.1983 in Köln
Anschrift:	Gliesmaroder Str. 14, 38106 Braunschweig Tel.: 0176 / 61731273 E-mail: s.quester@tu-bs.de
Familienstand:	ledig

Schulbildung

1989 – 1995	Besuch der Grundschule KGS Freiligrathstraße
1996 – 2002	Besuch des Gymnasiums Hildegard-von-Binden-Schule Abschluss Abitur (Durchschnittsnote: 2,3)

Ausbildung

2002 – 2008	Studium der Biologie an der Universität zu Köln
12.03.2008	Abschluss Diplom (Gesamtnote: sehr gut)
ab 15.4.2008	Promotion im Fach Biotechnologie an der Technischen Universität Braunschweig